# Genotype Networks: Convergent Evolution, Population Size and Adaptation

---

**Dissertation**

zur
Erlangung der naturwissenschaftlichen Doktorwürde
(Dr. sc. nat.)

vorgelegt der
Mathematisch-naturwissenschaftlichen Fakultät
der
Universität Zürich

von
**Reza Ali Rezaee Vahdati**
aus
dem Iran

**Promotionskommission**

Prof. Dr. Andreas Wagner (Vorsitz)
Prof. Dr. Lukas Keller
Prof. Dr. Kentaro Shimizu

Zürich, 2017

*"Science may be described as the art of systematic oversimplification."*

Karl Popper

# Abstract

Network science provides means to study complex systems in a way that is not possible by reducing the systems to their constituent elements. Genotype networks make it possible to better understand the behavior and phenotype of cells as complex biological systems. Here, I present the use of three genotype networks to answer different biological questions. In Chapter 2, I construct a haploid genotype (haplotype) network for each one of 12,235 human protein coding genes using data from the 1,000 genomes project. Notwithstanding their shared evolutionary history, these networks show widely different structures, indicating different patterns of variation. I further analyzed those genes that have more cycles in their associated network than expected by chance alone (42 genes). The occurrence of these cycles can be explained by parallel or convergent evolution at the sequence level. To further test this hypothesis, I analyzed the effect of positive, purifying, and balancing selection, as well as constrained mutations, on the occurrence of these cycles. Constrained evolution and purifying selection potentially had a major role in the origin of these cycles. Additionally, I found evidence of positive selection on 21 genes. Balancing selection, however, had at most a small role in bringing forth the excess of cycles. In Chapters 3 and 4, I simulate the evolution of populations on two different genotype networks to determine the effect of population size and mutation rate on evolutionary adaptation. In Chapter 3, I construct genotype networks from arbitrary RNA sequences, in which RNA sequences correspond to network nodes and computationally predicted RNA secondary structures are used to determine the "fitness" of the nodes. In Chapter 4, I construct 957 genotype networks from empirically determined binding affinities of transcription factors to eight-nucleotide-long DNA sequences. Using both types of networks, I observe that small populations, even in the most rugged landscapes, have no adaptive advantage over large populations. However, population size, even at constant population mutation rate, can dramatically affect evolutionary properties of populations, such as the rate of sequence exploration and population diversity.

# Abstrakt

Die Netzwerkanalyse liefert Mittel, um komplexe Systeme in einer Weise zu studieren, die durch die Reduzierung der Systeme auf ihre Bestandteile

nicht möglich ist. Genotyp-Netzwerke ermöglichen es, das Verhalten und den Phänotyp von Zellen als komplexe biologische Systeme besser zu verstehen. Hier analysiere ich drei Genotyp-Netzwerke, um verschiedene biologische Fragen zu beantworten. In Kapitel 2 konstruiere ich anhand der Daten aus dem 1000 Genome-Projekt haploide Genotyp-(Haplotyp)-Netzwerke für 12235 menschliche Gene. Ungeachtet ihrer gemeinsamen evolutionären Geschichte weisen diese Netzwerke sehr unterschiedliche Strukturen auf, die auf verschiedene Muster der Variation hinweisen. Weiters analysiere ich diejenigen 42 Gene, die Zyklen in ihren Netzwerken haben welche nicht allein durch zufällige oder neutrale Prozesse erklärt werden können. Das Auftreten dieser Zyklen kann durch parallele oder konvergente Evolution auf der Sequenzebene erklärt werden. Um diese Hypothese weiter zu prüfen, analysiere ich die Auswirkung der positiven, reinigenden und ausgleichenden Selektion, sowie die Auswirkung eingeschränkter Mutationen in der Entstehung dieser Zyklen. Eingeschränkte Evolution und reinigende Selektion spielten möglicherweise eine wichtige Rolle in der Entstehung der Zyklen. Darüber hinaus fand ich bei 21 Genen Hinweise auf positive Selektion. Im Gegensatz dazu spielte die ausgleichende Selektion bei der Entstehung der Zyklen höchstens eine kleine Rolle. In den Kapiteln 3 und 4 simuliere ich die Evolution von Populationen auf zwei unterschiedlichen Genotyp-Netzwerken, um die Wirkung der Populationsgröße und Mutationsrate in der Adaption von Populationen zu bestimmen. In Kapitel 3 konstruiere ich Genotyp-Netzwerke aus zufällig ausgewählten RNA-Sequenzen. Die Knoten dieser Netzwerke entsprechen RNA-Sequenzen, wobei die „Fitness“ jeder dieser Sequenzen sich aus deren rechnerisch prognostizierten RNA-Sekundärstruktur ergibt. In Kapitel 4 konstruiere ich 957 Genotyp-Netzwerke aus empirisch bestimmten Bindungsaffinitäten von Transkriptionsfaktoren. Bei beiden Arten von Netzwerken beobachte ich, dass kleine Populationen, sogar bei äußerst zerklüfteten Landschaften, keinen adaptiven Vorteil gegenüber großen Populationen haben. Jedoch kann die Populationsgröße auch bei einer konstanten Populationsmutationsrate, das evolutionäre Verhalten von Populationen, wie beispielsweise die Geschwindigkeit der Sequenzexploration und die Anhäufung genetischer Diversität erheblich beeinflussen.

# *Acknowledgements*

First and foremost, I am heartily grateful to my supervisor Andreas for giving me the chance to work in close collaboration with him and to learn about quality scientific work and vision. He gave me great scientific freedom and supported me in all the moments of success and failure. I never ceased to be impressed by how dedicated and precise he is. I would like to show my gratitude to my PhD committee members Lukas and Kentaro whose comments greatly helped me finish my projects.

It was an honor for me to get to know all of my friends and colleagues: Giovanni, Jose, Mariana, Manuel, Tugce, Kasia, Sinisa, Athena, Yolanda, Rzgar, Aditya, Debbie, Heidi, Macarena, Jia, Riddhiman, and Peter. I am especially indebted to Kathleen for all the discussions we had and for her support throughout my PhD. I thank Josh for helping me write a chapter of my dissertation; I learned a lot, Josh. I will always remember the exciting conversations I had with Charles, Fahad, and Magdalena over coffee.

I have been lucky to know Carlos Melian. I have learned a lot from his character and vision. I will not forget his help to finish my PhD. I would also like to thank Annette's administrial support.

I would like to thank my friends who made life joyful: Mohammad, Fabienne, Navi, Mehdi, Meysam, and Raha.

Last but not least, I owe deep gratitude to my family: to my parents for their ceaseless emotional and intellecutal support, and to my brother and sister without whome I would not have started this PhD. My love, Peymaneh, was beside me through all the pleasure and frustration. I am profoundly thankful to you.

# Contents

*To my Peymaneh*

# Chapter 1

# Introduction

## 1.1 Network science, a brief history

How can the Argentine ants in southern Europe control and organize a supercolony comprising of billions of worker ants distributed over 6,000 kilometers [85]? How do auklets flock and golden shiners school in such an organized manner? How do proteins in a cell interact to provide growth and maintenance? These examples, and many more biological (e.g. metabolic networks), technological (e.g. the world wide web), and social (e.g. large organizations) examples represent complex systems, where we cannot predict the collective behavior of a system from knowing the behavior of their constituent parts [13, 171].

Network science provides tools to study complex systems. All the systems mentioned above, despite their remarkable differences, can be represented with networks. A network consists of nodes (vertices) and links (edges). For example, each ant in a colony can be represented by a node, and communications between two ants by a link, which collectively, constitute a network of communications between ants in a colony. In a network of protein interactions, the proteins in a cell are the nodes of the network and interactions between any two proteins are the links between the nodes. Notwithstanding the differences among complex systems in size and types of objects representing the nodes and the links, and even in the processes that create these systems, a network representation can help uncover similarities between them (Figure 1.1). The same principles govern the behavior of different complex systems, and network science has discovered many such principles. In the following chapters of this thesis, I will use different kinds of networks to answer biological questions, such as how to detect convergent evolution among

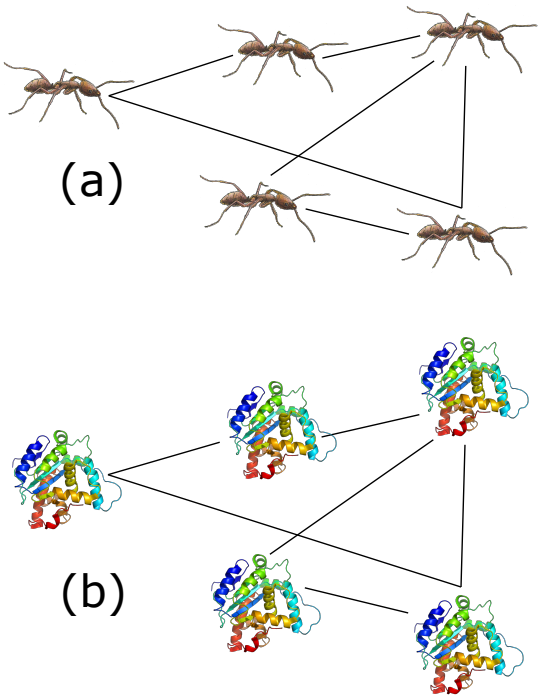


FIGURE 1.1: Two example networks. **(a)** A network of interactions between ants, where each ant is a node and communications between ants constitute the links of the network. **(b)** A network with the same structure as (a), but representing protein-protein interactions.

genes from human genetic variation data (Chapter 2), and how to determine the role of population size in genotype networks constructed from RNA secondary structures (Chapter 3) and transcription factor binding sites (Chapter 4).

**A short history of network science** Network science takes its mathematical formalism from graph theory, a subfield of mathematics [24]. Graph theory, unlike many disciplines, can trace back its roots to a certain point in time. Leonhard Euler, a Swiss-born mathematician, wrote a paper on the problem of the seven bridges of Königsberg in 1735 [66], which marks the dawn of the graph theory. Königsberg, the old name of the capital of Eastern Prussia, now called Kaliningrad and located in Russia, was a major trading center at the time. The river Pregel passed through the city, and its two branches created two islands in the city (Figure 1.2a). It was a prosperous city with trading ships filling the Pregel. The residents had built seven bridges connecting the mainland to the two islands. The positioning of the bridges raised a question which remained unsolved until 1735: How can one walk across all seven bridges and not cross any bridge twice? In 1735, Euler wrote a paper (which

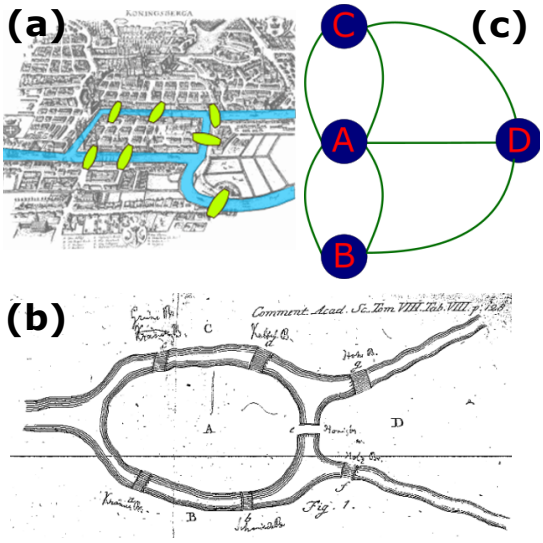


FIGURE 1.2: Königsberg's bridges and Euler's solution. **(a)** A historical map of Königsberg in 1651 with its river Pregel and its seven bridges highlighted (author: Merian-Erben, from Wikimedia commons). **(b)** Figure 1 from Euler's paper on solving the seven bridges problem [66]. Euler assigned a letter to each parcel of land and connected the letters with links where there was a bridge. He made a figure similar to **(c)**. This simplification allowed him to use a graph to solve this problem.

he later presented in 1736 and published in 1741) [66], that proved there is no answer to the bridges problem. This paper is regarded as the first one in graph theory. Euler simplified the problem by assigning a letter to each parcel of land separated by the river from the other and connected the letters by lines where bridges connected the parcels (Figure 1.2b and c). This presentation helped him eliminate unnecessary information and focus on the essential information, i.e. links and nodes. He observed that to cross all links (bridges) only once, one needs to enter a node and exit it an equal number of times, except for the nodes that mark the start and end of the travel. Thus, if a network has more than two nodes with an odd number of links, one cannot cross all links and never cross a link more than once. The network of bridges has four nodes, all of which have an odd number of links. Therefore, the problem is unsolvable.

Although the roots of network science can be traced back to the beginings of graph theory, it has flourished only in the 21st century [13]. In addition to using mathematical formalism from graph theory, network science uses techniques to handle randomness and noisy data from statistical physics. The reason that network science has flourished only recently is twofold. First, we have only recently been able to store, share and map large amounts of data.

Multiple ongoing projects create and analyze large-scale networks. Examples include the CAIDA project [241], which maps the Internet, databases of protein-protein interactions across the tree of life [236], or the Connectome project, which aims to map all neural interactions in the human brain [232]. Data availability is central to network science. Second, only after various datasets had become available were scientists able to analyze them, and discover that many complex systems, despite their profound differences, are governed by the same principles [13].

### 1.1.1 Basic definitions in graph theory

I provide some basic definitions and properties of graphs, which we will encounter throughout this thesis.

**Directed network.** An edge is directed if it has a direction associated with it, and a network whose edges are all directed is a directed network. For example, a predator-prey network that indicates which organisms in a community prey on other organisms, is a directed network. In contrast, a network of protein-protein interactions is undirected. A network may have both directed and undirected edges. For example, a metabolic network, which represents the metabolic reactions in a cell, can have reactions that are bidirectional (reversible) and reactions that are unidirectional (irreversible).

**Degree and degree distribution.** The degree of a node is the total number of its neighbors. In networks built from real data, not all nodes may be connected; indeed, the total number of links in such networks is wholly much smaller than the maximally possible number of links [13]. The probability that a randomly chosen node in a network has $k$ neighbors defines the degree distribution of a network. The degree distribution is an important characteristic of networks, because its form can help understand many network phenomena, such as robustness (the ability of a network to maintain its function despite perturbations).

**Distance.** Distance in networks is different from distance in the physical world. The distance between two nodes in a network is the shortest number of links that connects them. The series of links that takes us from one node

to the other is called a **path**. There may be several paths between two nodes, and the shortest path is the one with the fewest links. A path that starts from one node and ends at the same node, while it does not visit other nodes more than once, is called a **cycle**.

**Diameter.** The diameter of a network is the longest path among all the shortest paths between every pair of nodes.

**Connected network.** A network is connected if there is a path between any two nodes in the network. A group of nodes that can be reached from one another form a **component** of a network.

**Clustering coefficient.** The clustering coefficient of a node specifies how connected are its neighbors to one another [13]. The more links between the neighbors of a node exist, the higher is the clustering coefficient of the node.

## 1.2 Genotype networks

Genotype networks consist of a set of genotypes (nodes) with the same phenotype. A link connects two genotypes with a minimal mutational difference. Examples of a minimal mutational difference include a single nucleotide difference between DNA/RNA seqeuences, an amino acid difference between proteins, or deletion/addition of a reaction in metabolic networks. Talking about protein spaces, John Maynard Smith hinted to the concept of genotype networks for the first time in a paper in 1970 [165]. He referred to a set of proteins, some functional and some non-functional, that are connected by single amino acid mutations. Through these single amino acid mutations, one protein may change into another protein, while all the intermediates are also functional proteins. His concept of network includes genotypes that have any phenotype in the same network, whereas we limit the definition of genotype networks to genotypes with a similar phenotype. A breakthrough in constructing genotype networks started with Lipman and Wilbur in 1991 [145] by providing a genotype-phenotype map of lattice model of proteins.

Three classes of genotype networks have been studies most extensively: molecular networks of RNA [225], DNA [50, 211, 251], and proteins [146]; gene regulatory networks [8]; and metabolic networks [208].

**Molecular genotype networks** In RNA genotype networks, the genotypes are ribonucleotide sequences, and phenotypes refer to their fold or biological functions. In DNA genotype networks, the genotypes are nucleotide sequences of genes and phenotypes refer to their functions. In protein networks, genotypes correspond to amino acid sequences and phenotypes correspond to protein folding or biological function. In all molecular genotype networks, a link connects two genotypes if they differ by a single nucleotide/amino acid mutation or a small insertion/deletion (indel).

**Genotype networks of regulatory networks** A gene regulatory network consists of DNA sequences that determine regulatory interactions among regulatory proteins. By regulation, I mean any activity that affects the expression of a gene product. This activity can include changes in enzyme efficacy; signaling protein activities; mRNA abundance; and rates of gene transcription, which is the most common means of gene regulation [255]. Transcriptional regulation is mediated by binding of proteins (transcription factors) to DNA sequences (transcription factor binding sites) near genes, which increases or decreases the rate of transcription from DNA to RNA. Transcription factors regulate gene expression by inhibiting/facilitating the recognition of a gene by RNA polymerase. The genotype of a gene regulatory network comprises the DNA sequences encoding its regulatory interactions. Its phenotype comprises its gene expression patterns or concentrations of regulatory molecules. A link connects any two regulatory networks that differ in a regulatory interaction.

**Genotype networks in metabolism** Genomes of different organisms encode enzymes that catalyze the biochemical reactions inside cells. Such reactions make it possible to convert food molecules into smaller essential molecules, produce energy and cellular building blocks, and execute many other cellular functions. With our current knowledge of enzymes, we can build networks that represent the flow of material and energy inside a cell through biochemical reactions. In such metabolic networks, a genotype corresponds the DNA that encodes enzymes catalyzing metabolic reactions, and a phenotype refers

to the ability of a cell to synthesize biomass and produce energy from a given set of nutrients. A link connects any two metabolic networks that differ by a single reaction.

Note that by my definition, the genotypes in a genotype network have a single phenotype. The genotype space, which is the network of all possible genotypes regardless of their phenotype, is much larger, and can include many genotype networks.

### 1.2.1 Properties of genotype networks

I will now summarize a few important features of genotype networks, which are common among different networks, e.g. metabolic and gene regulatory networks.

**There are more genotypes than phenotypes.** An essential property of a genotype network is that for each phenotype there are many genotypes. For example, a protein of length 100 has $20^{100} \approx 10^{130}$ possible genotypes (because there are 20 different amino acids), but there are only about $10^4$ protein tertiary structures [255]. As another example, RNA sequences of length $N$ have $4^N$ possible genotypes, whereas there are only about $1.8^N$ secondary structures [255]. Similarly, the number of genotypes for gene regulatory and metabolic network is orders of magnitude larger than the number of phenotypes. The existence of many genotypes for the same phenotype is a requirement for existence of **neutral networks**. The term was first coined by Schuster et al. [225]. They considered a network neutral when changing the genotype does not lead to a different phenotype. This definition of neutrality is different from what is used in evolutionary biology, which is any mutational change that does not change the fitness of an organism. Neutral networks are important for helping populations find novel phenotypes [254].

**Genotype networks extend far through genotype space** A single genotype network often constitutes a small fraction of genotype space. Nevertheless, such a genotype network can extend far across genotype space, even as long as diameter of genotype space. For example, ref. [255] showed that if we randomly choose two gene regulatory networks with the same phenotype, they differ on average in 80% of their interactions. Ref. [40] compared the

distance between many pairs of randomly sampled genotypes in a gene regulatory network and found that this distance can be as large as the diameter of genotype space.

**Different genotype networks are interwoven** The minimum number of mutational steps required to reach from a random node in one genotype network to a node in arbitrary different genotype network is often a small fraction of the genotype space [255]. For example, ref. [164] estimated the distance between two arbitrary metabolic networks. To that end, they chose 1,000 random pairs of metabolic network genotypes. The genotypes in each pair had different randomly chosen phenotypes. They then counted the smallest number of random mutational steps it takes for one genotype to approach the other, without changing its phenotype on the way. The average distance between 1,000 phenotype pairs was one tenth of the maximally possible distance. Sequences in a radius of one tenth of the diameter of a genotype space comprise only a tiny fraction of all sequences. This shows genotype networks are not only spread across genotype space, but they are well intermixed and reachable to one another.

## 1.3 Convergent evolution

Convergent or parallel evolution, which is the focus of the next chapter, occurs when similar phenotypes or genotypes evolve in different lineages in response to similar selective pressures. The difference between convergent and parallel evolution is not sharply defined, but parallel evolution is often considered to occur in evolutionarily closely related lineages, whereas convergence occurs in evolutionarily more distant lineages [152]. Convergence is less likely to happen by chance alone than parallel evolution [181].

Examples of convergence at the morphological phenotypic level include the similar body shapes of sharks (fishes) and dolphins (mammals), or the similar wing shapes of birds and bats. An example of convergence at the protein level includes hemocyanin, a respiratory protein that functions similar to hemoglobin. It is found in taxa living in deep sea waters, such as arthropods. Hemocyanin contains copper instead of hemoglobin's iron, and copper gives blood a blue color. Despite the similar functions of hemocyanin and hemoglobin, they have little sequence similarity. Another example is

the repeated evolution of enzymes cleaving peptide bonds, which include sulfhydryl peptidases, metallopeptidases, aspartyl peptidases, serin peptidases, etc. [57, 213]. We know of several different molecules that show convergence [32, 52, 133, 147, 199, 250, 276]. Examples include peptide-binding regions of human and mouse class Ib genes in the major histocompatibility complex (MHC) [276], and the motor protein Prestin, vital to mammal's auditory system, of echolocating bats and echolocating dolphins [147].

Strong selection is generally considered the main driver of convergence. A famous example of this process is the repeated reduction of the number of armor plates of skeleton in freshwater threespine sticklebacks compared with marine sticklebacks [43]. Other mechanisms, however, can also drive convergence. Different mutation rates at different loci in the genome, a population's distance from an optimal genotype, clonal interference, and the size of a population can increase the probability of convergence [12]. For example, as sizes of populations increase, as least up to a certain limit [235], they are more likely to follow a similar mutational trajectory because they simply produce more beneficial mutations, and are thus more likely to find large effect mutations [12]. Populations closer to a fitness peak have fewer beneficial mutations available to them, and are more likely to experience the same ones. Clonal interference can also increase the chance of convergence by increasing the probability that large effect mutations become fixed [12].

A remarkable phenomenon about phenotypic convergence is that when a specific trait is under positive selection, often the same genes and mutations are used for producing the trait. An example is the evolution of fluoroquinolone resistance in *Pseudomonas aeruginosa*, where only a few genes [265] out of over a hundred possible genes [27] contribute to resistance evolution. Three causes can explain this phenomenon. First, some genes may be located in regions of the genome with high mutation rates, or they may simply be larger than other genes, thus experiencing more mutations. Second, genes that affect multiple traits are less likely to be the subject of adaptive evolution in a single trait, because it can be more difficult to preserve the other functions of the gene. A potential example of a gene affecting few traits is the *Mc1r* gene, a gene that affects pigmentation across vertebrates. Third, some genes may show genetic variation that is pre-adapted to new environmental conditions [152].

## 1.4 Human genetic variation

In Chapter 2, I will analyze genotype networks from human populations to detect convergent genotypic evolution. I will thus briefly describe nature of genetic variation, its causes, and some methods to detect it.

### 1.4.1 A short review of human evolutionary history

Hominins (including all extant and extinct human species) separated from other apes 5-7 million years ago in East Africa [73]. Modern humans (*Homo sapiens*) belong to the genus *Homo* from the hominin clade, which also includes our extinct ancestors and relatives, most notably *H. erectus* and *H. neanderthalensis*. The genus *Homo* diverged about 2.5 million years ago from other hominins (Australopithecines) by evolving a bigger brain ($640cm^3$ versus previous $500cm^3$) and a smaller jaw, which coincides with the first archeological evidence of stone tools [73].

The first genetic analyses to uncover the history of modern humans used mitochondrial DNA and were published in the 1980s [30]. The human mitochondrial genome, a circular DNA with 16,569 base pairs and 37 genes, is passed only from mothers to their offsprings [6]. This makes it a good candidate for studying evolutionary history because recombination between paternal and maternal lineages can otherwise obscure evolutionary events. An early mitochondrial DNA study [30] led to the mitochondrial Eve hypothesis: if we trace the lineage of all human mitochondria, it goes back to a female living in Africa in about 200,000 years ago. The female probably lived in a population of around 10,000 individuals. Note that the mitochondrial Eve hypothesis does not state that there was a single female living at that time, but since only the maternal lineage contributes to the inheritance of the mitochondrial genome, a population loses a fraction of its mitochondrial genetic variation every generation (some females only have male offspring, and their mitochondrial genome does not contribute to the next generation) [73]. This loss of variation over many generations in the human population is responsible for the fact that all current mitochondrial genetic variation have the same ancestor 200ka. The mitochondrial Eve hypothesis and the single-origin of humans in Africa are supported by further mitochondrial and nuclear genome analyses [73]. At around 100ka [90] humans began to emigrate out of africa in various waves, and eventually occupied all continents [158].

### 1.4.2 Human effective population size

The effective size of a population is an important parameter that affects the amount of genetic variation. For humans, a number of approximately 10,000 individuals is often used as the effective population size [277]. Effective population size, first introduced by Sewall Wright in 1931 [270], is the size at which a population experiences the same amount of evolutionary change as an idealised population. An idealised population is a theoretical concept in evolutionary biology that describes a population with certain characteristics, such as unchanging population size, the absence of migration, random mating, etc. When a population size changes over time, one can use the harmonic mean of the population over time to calculate its effective size [270]. This is the reason that human effective population size, despite its current large size, is so small: humans have had a long history of small populations which strongly affects the harmonic mean of population size to this day.

### 1.4.3 DNA sequencing techniques

DNA sequencing methods have deepened our knowledge about the evolutionary history of our and other species, and about fundamental principles governing evolutionary processes. The first complete genome of an organism, that of bacteriophage MS2, was published in 1976 [72]. As of now, genomes of 361 eukaryotes, 7,421 prokaryotes, and 7,142 viruses have been completely sequenced and published [116, 180]. I will present a brief review of popular sequencing methods and their advantages and limitations.

#### Sanger sequencing

In 1977 Frederick Sanger and his associates developed a sequencing method (Sanger sequencing or the chain termination method) [220], which became the most widely used sequencing method for several decades. The following materials are required for Sanger sequencing: a solution of target DNA fragments to be sequenced, DNA primers, DNA polymerase enzyme, deoxynucleoside triphosphates (dNTPs), and modified di-deoxynucleotide triphosphates (ddNTPs). the ddNTPs may be radioactively or fluorescently labeled. The process starts with a pool of target DNA. High temperature converts the double-stranded target DNA to single-stranded DNA. Next, primers anneal

to target DNA sequences at a lower temperature. Finally, DNA polymerase uses dNTPs or ddNTPs to replicate the target sequence. ddNTPs lack a 3'-OH group in their structure that terminates the extension process. The concentration of ddNTPs is much lower than that of dNTPs, which allows production of fragments with different lengths. If ddNTPs are radioactively labeled, four separate sequencing reactions are needed to perform DNA replication reactions, where each reaction contains only one of the ddNTPs and all other dNPTs. If ddNTPs are fluorescently labeled, each with a different color, a single sequencing reaction is enough. The fragments may be separated using gel electrophoresis, and the sequence of target DNA can be read from a gel. The order of bands after gel electrophoresis shows the sequence of the target DNA. Sanger sequencing can be used to sequence DNA fragments of about 1000bp [227]. It is a widely used sequencing techniques to date, and is employed in important research projects, such as the Human Genome Project [138]. Among the limitations of Sanger sequencing is that the initial 15-40 base pairs of the target sequence have low quality because of primer binding. Furthermore, Sanger sequencing can only detect substitutions and small insertion/deletions. Moreover, using Sanger sequencing one may not detect mosaic mutations, i.e. mutations occurring after fertilization and at different frequencies within tissues.

**Next-generation sequencing**

Next-generation sequencing (NGS) techniques provide the ability to sequence large amounts of DNA in a much shorter time and with a lower cost than on Sanger sequencing. NGS technologies can sequence millions of DNA fragments at the same time. There are multiple commercial approaches to next-generation sequencing that have different sequencing biochemistry and library generation. However, they can be classified as cyclic-array sequencing. These methods use the physical separation of DNA fragments, and several cycles of enzymatic reactions in sequence generation [111]. The first NGS technology was 454 pyrosequencing [161]. Other popular NGS technologies include the Illumina sequencing by synthesis [71], AB SOLiD [228], and HeliScope [95]. NGS can be used for whole genome sequencing, targeted sequencing of selected genomic regions, and de novo sequencing. A limitation of NGS technologies is their limited power to resolve regions of a genome with many repeats. Since sequences need to be assembled from many small fragments, repeats can create ambiguities in assembly. Depending on the

application, different NGS technologies may have advantages and disadvantages. The details of their procedures and limitations are beyond this introduction but can be found elsewhere [227].

Newer sequencing technologies do not need to amplify template DNA for sequencing, but rather use the template DNA directly. PacBio [197] and Nanopore [196] are the most notable examples of such sequencing technologies. An advantage of these technologies is that they can produce significantly longer reads, which can alleviate computational complexities of de novo genome assembly. For example, PacBio can produce thousands of sequence reads at an average length of exceeding 20,000 pb.

### 1.4.4 Structure of the human genome

The first draft of the human genome was published in 2001 [138], and the complete version was published in 2003 [106]. With these publication, our knowledge of the human genome architecture increased significantly. Most of the human genome resides in the nucleus (more than 3 billion base pairs). The nuclear genome is organized in 23 homologous chromosomal pairs (46 chromosomes), one from each paternal and maternal lineage. Of the 46 chromosomes, 44 are autosomal chromosomes, and two are sex chromosome. There are 20,267 protein-coding and 19,102 RNA coding genes in the human genome [116].

Eukaryotic genes are much more complicated than prokaryotic ones, because their sequence does not usually directly translate into an amino acid sequence. Eukaryotic protein-coding sequences (exons) are interspersed among introns that do not translate into proteins. There are on average 7.8 introns per gene in the human genome [219], many of which are also well-conserved [130]. Furthermore, complication occur when multiple translated regions in a gene overlap, which means that introns in one gene can be exons in another overlapping gene [25].

### 1.4.5 Human genetic variation

**Sources of genetic variation.** The primary source of genetic variation in a population is DNA mutation. The most common type of mutations is point mutations, mutations that result in substitution, insertion, or deletion of a

single nucleotide. The fraction of mutations that are neutral or non-neutral depends on the genetic background and on environmental changes that a population experiences [121, 149, 184, 229]. A secondary source of genetic variation is gene flow or migration. It occurs when individuals from one population migrate to another and interbreed with the resident population. The extent of the effect of gene flow in changing a population's allele frequencies (frequencies of genetic variations in a population) depends on the level of genetic similarity between two populations, the amount of migration, and whether migration is unidirectional or bidirectional. Humans have a history full of migrations and admixtures between populations.

Some evolutionary and genetic processes do not generate new genetic variation but can change the frequencies of different alleles. Genetic drift is one of them. It causes a loss of genetic genetic variation due to random sampling of alleles. It is stronger in smaller populations. Specifically, the fate of any allele whose fitness is less than the reciprocal of population size ($1/N$) is governed by genetic drift. Genetic drift always tends to reduce genetic variation over long timespans. Genetic recombination is another process through which genotype frequencies can change. Recombination is the combination of paternal and maternal chromosomes during meiosis, which generates new combination traits to be passed to offsprings. Natural selection, a force that explains adaptation and speciation, is another force that changes allele frequencies. Depending on the type of selection, variation may decrease (e.g. purifying selection) or increase (e.g. balancing selection).

**Statistics about human genetic variation.** A breakthrough in identifying human genetic variation came from the first phase of the 1,000 genomes project in 2010 [61] and from the project's completion in 2012 [168]. The 1,000 genomes project sequenced 1,092 individuals from 14 populations in four continents (Europe, East Asia, sub-Saharan Africa and the Americas). The genomes were sequenced by a combination of low-coverage whole genome sequencing, high-coverage exome sequencing, and SNP genotyping. The project used different sequencing technologies including ABI_SOLiD, Illumina, and 454 [168]. It identified most SNPs and indels in the human population, except rare SNPs (e.g. an estimated discovery of 50% of SNPs of frequencies about 0.001 and 98% of SNPs with a frequency of 0.01) and most short insertion/deletions (indels). Specifically, it identified a total of about 36.7 million SNPS and 1.38 million indels, with an average of 3.60 million

SNPs and 344,000 indels per individual. Genetic variants with a frequency $\geq 0.1$ is present in all major population ancestries of the project (European, African, East Asian), whereas 53% of rare variants (frequency $\leq 0.05$) are found in only one population. Parts of human evolutionary history reveal themselves in genetic variant frequencies. For example, low-frequency alleles at frequencies between 0.005 and 0.05 are about three times more common in populations with African ancestry. This points to bottlenecks in the recent history of non-African populations. Also a high number of rare variants (frequency $< 0.005$) in all human populations points to recent population expansions [168].

## 1.5 The effect of population size and mutation rate on evolutionary adaptation

Many processes affect the adaptation of organisms. Examples include the size of a population, the mutation rate,the distribution of fitness effects, the mode reproduction, migration, the age structure of a population, and the age of reproduction. Understanding how these processes affect genetic variation and how they interact with one another can help elucidate sources of genetic diversity and predict future evolutionary changes. In this section, I will review the effects of the size of a population and its mutation rate on adaptation, because they are relevant for subsequent chapters.

When population geneticists discuss the effect of population size on the evolution of organisms, they refer to effective population size. As defined earlier, effective population size is the size of an idealised population in which genetic drift has the same effect as in a given study population. The effective size of almost all natural populations is below their census sizes [75]. Several processes cause this difference. Among them are: unequal sex ratios, large variation in the number of offspring, inbreeding and non-random mating, age structuring of populations, bottlenecks and changes in population size [35].

The effective size of a population affects several other quantities such as the rate of nucleotide substitutions [267], genome size [99], genome complexity

[156], and the mutation rate [112]. The rate of beneficial nucleotide substitutions has been used as a measure of adaptation rate [139]. Therefore, explaining the effect of population size on the substitution rate of mutations has been the subject of many studies. One can group mutations into five categories based on their effect on fitness [190], and I review the effect of population size on each category.

The first category of mutation encompasses is neutral mutations, those that do not have any fitness effects. The number of neutral mutations produced in each generation in a population equals the number of individuals in the population ($N$) times the genomic neutral mutation rate ($\mu_n$). Each of these mutations has a chance of being fixed relative to the inverse of the population size $1/N$. Multiplying these two values, we get one of the most significant results in population genetics $N\mu_n \times 1/N = \mu_n$, that is, the substitution rate of neutral mutations equals their mutation rate and is independent of population size [126]. This result is so robust that even complete genetic linkage to advantageous or deleterious mutations does not affect the substitution rate of neutral mutations [21].

Two other categories of mutations encompass slightly beneficial and slightly deleterious mutations. Population size is important in affecting the fate of these mutations. Because genetic drift determines the fate of any mutation whose fitness effect $s$ is less than reciprocal of population size $1/N$, genetic drift is the sole determinant of fewer mutations in large populations. In addition, the number of mutations that are neutral in a small population and non-neutral in a large population depends on the distribution of mutational effects. This distribution differs among species [121, 149, 184] and changes as a population adapts to different environments [229]. Genetic linkage can reduce the substitution rate of slightly beneficial mutations, and increase that of slightly deleterious mutations [83].

Strongly beneficial and strongly deleterious mutations are the two final categories of mutations. Their fixation probability depends on the magnitude of their fitness effects [92]. Large populations produce more beneficial and deleterious mutations than small ones. However, selection becomes more effective in fixing beneficial mutations and purging deleterious mutations in large populations. Thus, more beneficial mutations and fewer deleterious mutations fix as population size increases [112, 139].

A basic prediction about the substitution rate of mutations in populations of

different sizes is that since most mutations are deleterious, and since small populations are more likely to fix deleterious mutations, substitution rates in small populations should be greater than those in large populations [190]. Several studies have indeed shown that the rate of nonsynonymous to synonymous substitutions $\omega$ increases in small populations. For example, animal species restricted to islands compared with their mainland relatives, endosymbiotic bacteria compared with free-living ones, and hominids compared with other mammals with higher $N_e$, all show increased values of $\omega$ [132, 172, 268, 269].

Clonal interference is a process in which two or more different beneficial mutations compete for fixation [81]. It is more likely to occur when the population mutation rate $N\mu$ is large. Therefore, the adaptation rate of large populations can be slowed down by clonal interference. However, the occurrence of additional beneficial mutations in a large population can reduce the negative impact of clonal interference on adaptation [53]. Clonal interference is only relevant when there is no recombination between alleles, such as in clonally evolving bacteria, or in regions of the genome of sexually reproducing organisms where recombination rates are very low.

## 1.6 Prediction of RNA secondary structures

In Chapter 3, I will analyze the effect of population size and mutation rate using genotype networks of RNA sequences and their secondary structures. Here, I will briefly describe the functions and structures of RNA, as well as the history and the importance of computational prediction of RNA secondary structures, together with commonly used approaches for this purpose.

Ribonucleic acid (RNA) is a heteropolymer of ribonucleotides guanine (G), uracil (U), adenine (A), and cytosine (C), which has been a focus of research in past years due to its broad functions inside cells [204, 243]. Its best-known function is that of an information intermediate between DNA and proteins (mRNA), transferring amino acids for the construction of proteins (tRNA), and linking amino acids to form proteins (rRNA). RNA molecules, however, have many other functions as well. There are different classes of noncoding RNAs (those that are not translated into proteins) that perform enzymatic or gene regulatory functions. Examples include small nuclear RNAs (snRNAs),

which regulate gene expression by affecting RNA splicing; microRNAs (miRNAs), which repress mRNA translation; small interfering RNAs (siRNAs), which inhibit transcription; riboswitches, which regulate gene expression in response to environmental stimuli; and catalytic RNAs (ribozymes), which can function independently of protein molecules, and perform tasks in a variety of processes, such as replication, mRNA preprocessing, and mRNA splicing [42].

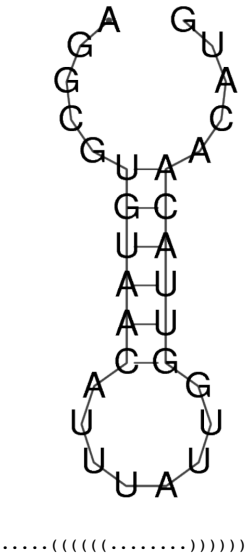


FIGURE 1.3: **Secondary structure of an RNA sequence.** The figure shows two representations of the predicted secondary structure of Z71666, a small nucleolar RNA (snoRNA) in *Saccharomyces cerevisiae* : on top, the sequence is shown with its covalent bonds between adjacent nucleotides, and with its hydrogen bonds between non-neighboring pairs. The bottom shows the dot-parenthesis representation of the secondary structure, where matching parenthesis represent hydrogen bonds between nucleotides, and dots represent unpaired nucleotides. This is the minimum free energy structure of Z71666, as predicted by the `fold` function in the ViennaRNA package [150]. It has a minimum free energy of $-4.90 kcal/mol$, and the sequence spends 0.35 of its time in this structure.

RNA molecules have a secondary (2D) and a tertiary (3D) structure. The secondary structure of RNA is a planar structure based on complementary base-pairing of nucleotides (A-U/U-A, C-G/G-C, and G-U/U-G). The tertiary structure of RNAs involves non-standard base-pairing, pseudoknots, and bivalent ions that bring together the elements of the secondary structure. RNA secondary structures have been studied for decades, both empirically and theoretically, and can be predicted with a sensitivity of more than 70% [150, 215]. However, it is more challenging to predict the tertiary structures of

RNAs [58, 258]. Nonetheless, predicting the secondary structure of RNAs can help predict their function because it is a prerequisite for properly formed tertiary structure [204]. Furthermore, RNA secondary structure is more stable than tertiary structure [242]. In addition, secondary structure can be closely linked to function on its own. For example, in mRNAs, the secondary structure determines the half-life of the molecule [160]. Secondary structure prediction has proven useful for studying genotype-phenotype maps [255].

### 1.6.1 Algorithms to predict RNA secondary structure

The first attempts to computationally predict RNA secondary structure were made in 1978, and aimed to solve the maximum circular matching problem [187]. This approach seeks to find a secondary structure with a maximal number of base pairs. In 1981, Zuker and Stiegler [283] introduced the first dynamic programming algorithm that constructs the secondary structure of an RNA molecule by minimizing its free energy. This approach still forms the basis of most widely used modern methods [150, 162, 283]. Such energy minimizations are based on empirical values of stacking energies, i.e. free energies of base pairs adjacent to one another, hairpins, internal loops, and other structural elements. Empiricaly measured energy values have been regularly updated as more data became available [163]. The accuracies of predictions have been further increased using statistical learning [7]. In 1999, Stefan Wuchty and collaborators published a paper that demonstrates how to calculate all suboptimal secondary structures of an RNA sequence within a given energy interval above the minimum free energy [273]. Calculating suboptimal secondary structures provides an opportunity for studying plasticity in genotype-phenotype maps [253]. There are other algorithms that do not measure thermodynamic parameters directly to predict RNA secondary structures, for example, machine learning approaches that use stochastic context-free grammars (SCFG) [62, 218].

## 1.7 Microarray technology

We cannot predict the collective behavior of a complex system only from knowing its components' behavior. A solution to understanding biological systems, especially cellular systems, is to identify and map the interactions

between components of a cellular network and use theoretical tools of network science to analyze them. Human cells have about 20,000 protein-coding genes which encode an estimated 6.13 million proteins (including modified protein variants as a result of alternative splicing, single amino acid polymorphisms, and posttranslational modifications) [207]. Experiments on single molecules are inefficient for untangling the web of interactions among these cellular components. Microarray technology provides a precise and high-throughput technique for identifying biochemical activities of millions of biomolecules and interactions among them in a single experiment (e.g. up to $10^5$ interactions per $cm^2$ of a microarray [167]). Microarrays can help study gene expression patterns, antibody-antigen interactions, protein-DNA interactions, and many more aspects of cellular phenotypes [212]. In Chapter 4, I will use data produced from protein binding microarrays. To construct genotype networks and simulate populations on such networks. I will briefly describe the fundamentals of microarray technology and some of its prominent applications.

Before explaining microarray technology, I provide some definitions that will be useful for describing the technology. A library of biomolecules is a set of all molecules that differ from one another in a well-defined way. For example, a DNA library may contain all possible DNA sequences of length 10 or only a subset of such sequences. There are different classes of libraries that differ in their spatial organization. The simplest library is a mixture of randomly generated sequences [56], which can be used to find aptamers (short DNA or peptide molecules that bind to specific biomolecules) [120]. A second class comprises libraries whose elements are spatially separated through binding to different microscopic beads [78]. In such libraries, one does not have a priori knowledge about what sequence is bound to each bead, and this needs to be determined with further experiments. Finally, there are libraries whose elements are arrayed on a supporting surface such as a plastic or glass microscope slide, a film, or a semiconductor chip [63, 231]. In such an array, we know the exact location of each element of the library.

Once an array is ready, analytes can be applied to the array. A scanner detects and records any interaction between analytes and the array elements. Detection of interactions which needs to be ultrasensitive to detect single molecule reactions, can be mediated by labeling [65, 91], or through other methods, such as electrochemical methods [60, 272]. The large amounts of data produced by microarray experiments require the use of bioinformatics methods.

The most widely used types of microarrays are DNA microarrays [222]. They are used to detect gene expression patterns and to sequence mutations on a large scale. DNA microarrays consist of DNA oligonucleotides attached to a supporting surface. The sample to be analyzed can be DNA or RNA (converted to cDNA) that can hybridize with the array elements. Only highly complementary sequences remain attached to the array after a washing step. Using microarray, one can determine gene expression levels of thousands of genes. Other applications include, but are not limited to, comparative genomic hybridization, chromatin immunoprecipitation on chip, single nucleotide polymorphism detection, and alternative splicing detection.

DNA microarrays are provide only limited information about protein abundance and functions. One reason is that protein expression levels do not always correlate with mRNA levels [89, 159]. To solve this problem, protein microarrays were first introduced in 1983 [33]. In a protein array, proteins or nonpeptide aptamers are used as the elements mobilized on a surface. Protein microarrays can be employed to identify protein expression, or to identify protein–protein interactions, disease biomarkers and the DNA-binding specificity of protein variants [272].

Protein binding microarray (PBM) technology comprises another category of microarrays that provides rapid, high-throughput characterization of protein-DNA interactions in vitro [19, 20, 176]. Complex response of cells to environmental changes and changes in gene expression throughout development are mediated by transcription factors binding to DNA sequences. Transcription factors can activate or repress gene expression by promoting or inhibiting transcription of genes. DNA binding sites of transcription factors in eukaryotes are usually short (6–10 base pairs) [19]. PBMs provide a means of measuring binding affinity of transcription factors to all possible DNA binding sequences in a single experiment. This helps building full landscapes of transcription factor binding affinities, and provide insight into the regulatory functions of transcription factors. Moreover, by quantifying binding affinity of transcription factors to all possible sequences, one can detect binding differences in homologous proteins. In PBMs, transcription factors are expressed with epitope tags (sequences that can be recognized by antibodies), and then applied to arrays of double-stranded DNA arrays. A washing step removes any transcription factor that is not highly complementary to DNA

sequences. Fluorophore-conjugated antibodies are then applied to microarrays, which attach to epitope tags of transcription factors and help quantifying binding of transcription factors to DNA sequences [18]. A limitation of current PBMs is that they can only detect binding affinities of transcription factors with short motifs (less than 12bp) [19]. Prokaryotic transcription factors can bind to DNA sites 20bp or longer.

## 1.8 Thesis outline

In the following chapters, I will describe three research projects. Chapter 2 focuses on a network-based approach to detect convergent genotypic evolution from human genetic variation, and chapters 3 and 4 simulate the evolution of populations on two different genotype networks to answer how the size of populations and its interaction with mutation rate can affect adaptive evolution.

In Chapter 2, I use a novel approach to analyze genetic variation in human genes. I use single nucleotide polymorphism (SNP) data of 1,094 individuals from four continents (Europe, East Asia, sub-Saharan Africa and the Americas) [168]. For each human gene, I construct a genotype network from non-synonymous SNP data, i.e. each node of the network corresponds to a haploid genotype (haplotype) of the gene. A link connects two nodes that differ by a single nucleotide. Traditionally, phylogenetic trees have been used to represent such genetic variation data. However, trees cannot easily represent cycles, and thus cannot represent more complex evolutionary relationships than vertical descent. Using networks, I study small cycles of few variants in human haplotype networks. Convergent evolution, in which two different sequences evolve to the same sequence, can explain the occurrence of such cycles. I find and describe 48 genes, out of 12,235 genes, which have an excess of cycles that cannot be explained by chance alone.

In Chapter 3, I study the dynamics of population evolution on adaptive landscapes. Specifically, I investigate the effect of population size and mutation rate on the adaptation of populations. I construct genotype networks from all DNA sequences with ten nucleotides that fold into some secondary structure. Additionally, I create genotype networks from four biological sequences of length between 30-43 nucleotides and compare their evolutionary dynamics with that of the shorter sequences. This approach in determining the role of

population size on adaptation bridges theoretical and empirical approaches. Theoretical studies, due to limited knowledge about natural adaptive landscapes, have to make simplifying ad hoc assumptions about the structure of such landscapes. Empirical studies, due to technical limitations, do not provide sufficient knowledge about topologies of an adaptive landscape and about mutational trajectories on such a landscape. RNA secondary structure prediction provides biophysically motivated adaptive landscapes, where one does not have to make ad hoc assumptions about landscape structure, distribution of mutational effects, epistatic interactions between mutations, etc. I show that population size is an important evolutionary parameter even when population mutation rates are equal. Population size influences parameters such as population diversity, and the ability of a population to explore multiple sequences, which affect population adaptation.

In Chapter 4, I ask the same questions as Chapter 3, but I use empirical adaptive landscapes from transcription factor binding sites. I consider 957 landscapes, each describing the binding affinity of a transcription factor to all of its 8-nucleotide long DNA binding sites. These landscapes vary in their ruggedness, measured as the number of landscape peaks. I find a strong negative correlation between landscape ruggedness and a population's mean fitness after 1,000 generations of simulated evolution. Analyzing nine of these landscapes in more detail, I make similar observation as in Chapter 3 about the role of population size in adaptation.

# Chapter 2

# Parallel or convergent evolution in human population genomic data revealed by genotype networks

Ali R. Vahdati, Andreas Wagner

# *Abstract*

**Background:** Genotype networks are representations of genetic variation data that are complementary to phylogenetic trees. A genotype network is a graph whose nodes are genotypes (DNA sequences) with the same broadly defined phenotype. Two nodes are connected if they differ in some minimal way, e.g., in a single nucleotide.

**Results:** We analyze human genome variation data from the 1,000 genomes project, and construct haploid genotype (haplotype) networks for 12,235 protein coding genes. The structure of these networks varies widely among genes, indicating different patterns of variation despite a shared evolutionary history. We focus on those genes whose genotype networks show many cycles, which can indicate homoplasy, i.e., parallel or convergent evolution, on the sequence level.

**Conclusion:** For 42 genes, the observed number of cycles is so large that it cannot be explained by either chance homoplasy or recombination. When analyzing possible explanations, we discovered evidence for positive selection in 21 of these genes and, in addition, a potential role for constrained variation and purifying selection. Balancing selection plays at most a small role. The 42 genes with excess cycles are enriched in functions related to immunity and response to pathogens. Genotype networks are representations of genetic variation data that can help understand unusual patterns of genomic variation.

## 2.1 Background

The patterns and causes of genotypic variation in human genes have been a focus of great recent interest in evolutionary biology. Different processes such as natural selection, genetic recombination, genetic drift, demography, as well as physicochemical properties of cells, can influence this diversity. Various methods have been devised to represent and quantify genetic variation and to detect its causes [1, 38, 70, 77, 168, 169, 217, 233, 237, 244].

Here we use a novel approach based on genotype networks to represent and analyze genetic variation in human genes. Genotype networks are graphs that consist of nodes, which correspond to genotypes with the same phenotype, where sameness can be defined as narrowly as enzyme activity, or as broadly as viability. Nodes that differ in some minimal way from each other are adjacent, i.e., connected by a link in such a graph. The genotypes we consider are haploid genotypes (haplotypes) of human genes in a sample of the human population, and we call two genotypes adjacent if they differ in a single nucleotide. Genotype networks can be useful to address various evolutionary questions, such as how novel adaptations originate, and what role phenotypic robustness or plasticity play in adaptation [252]. In the past, they have been mostly built from computational models of genotype-phenotype maps [41, 145, 164, 225], but high-throughput genotyping allows us to build genotype networks from experimental data [50]. Representing such data in the form of a network makes the large analytical toolbox of graph theory available, which has been useful in fields as different as ecology and the social sciences [15, 170, 183, 193].

A common tool for interpreting relationships among individuals using genetic variation data is the phylogenetic tree, which shows the evolutionary relationship among a set of taxa, individuals, or genes that constitute the leaves of the tree. The common ancestors of these taxa form the interior nodes of such a tree. In a gene tree, these ancestors can be reconstructed with the help of probabilistic models of sequence evolution [131, 275, 279]. Phylogenetic trees are by definition *acyclic* graphs: They do not contain cycles – paths of links that start from a node, pass through other nodes, and return to the same node.

The acyclic nature of phylogenetic trees implies one major limitation of such

trees: They cannot easily accommodate evolutionary genealogies more complex than simple vertical descent with modification [107, 136, 173, 174]. Such genealogies can lead to reticulate networks of phylogenetic relationships. Thus, multiple mechanisms to create genetic diversity, such as hybridization, allopolyploidization, sexual reproduction, recombination, gene conversion, and homoplasy, which lead to mosaic patterns of relationships among nodes are not easily accommodated in tree-like structures. Genotype networks provide information complementary to phylogenetic trees that are not subject to this limitation, because they can accommodate cycles. We note that cycles in genotype networks can also occur due to chance alone, meaning that different genealogies can lead to identical cycles. Inferring the true genealogy in a cycle requires further analysis.

Figure 2.1 shows a short cycle in a hypothetical genotype network involving four DNA sequences. Links reflect adjacent genotypes that differ in a single nucleotide. Assume, for example, that genotype 1 is ancestral to the other genotypes, and different substitutions (A10T and A20G) produce genotypes 2 and 3 from it. Genotype 3 then experiences an additional A10T substitution that creates genotype 4. This mutational path leads to a closed cycle, where three of the four links reflect substitution event. The fourth edge is a consequence of the first three events, because they render genotype 2 adjacent to genotype 4. Similar scenarios can be developed if a genotype different from genotype 1 is ancestral. Regardless of this ancestor, cycles require sequence changes that render the descendants of one (or more) genotypes more similar rather than less similar. In other words, cycles require some form of homoplasy, i.e., parallel or convergent evolution [39, 86, 151, 152, 257]. More generally, homoplasy is said to exist when two lineages display the same genetic or phenotypic characters, even though this similarity has not arisen through common ancestry [39, 86, 151, 152, 257].

Homoplastic sequence evolution has been documented in a wide variety of molecules [32, 52, 133, 147, 199, 250, 276]. It can be caused by chance alone, which is expected to be rare in long evolving biopolymers with multiple kinds of monomers, because random mutations are more likely to cause such polymers to diverge than to converge. Mutational biases, strong selective constraints on sequence evolution [52], positive selection [32, 52, 133, 147, 199], or genetic recombination [155] can also cause homoplasy.

Here we construct haploid genotype networks for each of 12,235 genes in the human genome, based on single nucleotide variation data available for

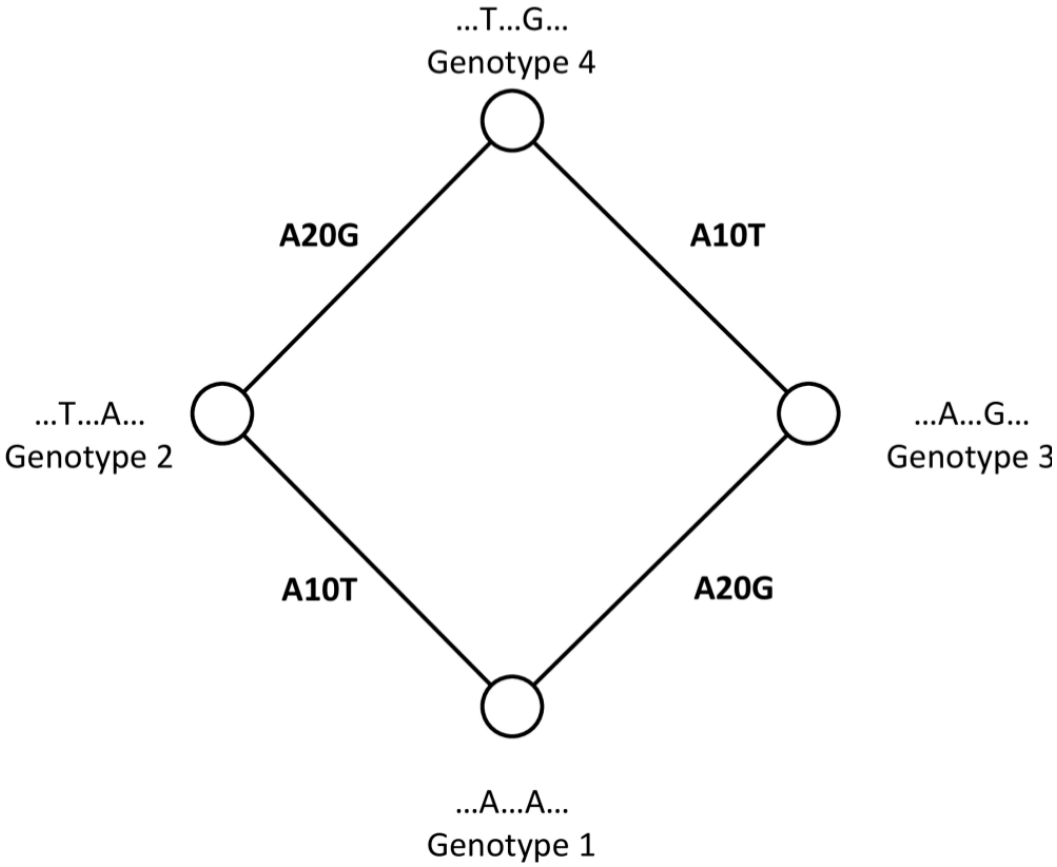


FIGURE 2.1: **A hypothetical example of a four-node cycle in a haplotype network.** The example indicates a hypothetical DNA sequence where two nucleotide changes occur at position 10 and 20. Circles (nodes) correspond to genotypes. A link connects two nodes if they differ by a single mutation. Lettering next to each node indicates the nucleotides at which two genotypes differ. Edge labels show changes required to create a genotype from its neighbor, e.g., "A20G" indicates a change from *A* to *G* at position 20 of the hypothetical sequence. In the example genotype 1, through mutation at positions 10 and 20 creates genotypes 2 and 3. Then, either genotype 2 mutates at position 20 from *A* to *G*, or genotype 3 mutates at position 10 from *A* to *T*, or both of these mutations happen together, and create genotype 4.

1,092 individuals from four continents [168].We analyze short cycles up to length eight in these networks, and find that the haploid genotype (haplotype) networks of 42 genes show a significant excess of cycles that cannot be explained by chance alone. After having excluded recombination as a prominent cause of these cycles, we focus on positive selection as a possible cause, and present evidence that in at least some of these genes positive selection may help explain the existence of cycles.

## 2.2 Results

### 2.2.1 Constructing and characterizing haplotype networks

To construct genotype networks for 1,092 human individuals, we used haploid genotypes (haplotypes) with single nucleotide variation data available from the 1,000 genomes project [168]. Thus, our genotype networks are haplotype networks, and from now on, we use the term haplotype network instead of genotype network. For each human gene, we constructed one haplotype network. Two principal definitions of such a network are germane for this paper. By the first definition, a haplotype network of a human gene is a graph whose nodes correspond to protein-coding DNA sequences of the gene in different individuals. Two nodes (sequences) are adjacent if they differ in a single base pair (i.e., by either a synonymous or non-synonymous change). By the second definition, two nodes are adjacent if their coding sequences differ by a single non-synonymous (amino acid replacement) change. The second kind of network can thus also be viewed as a network of proteins or amino acid sequences, in which neighboring proteins differ in a single amino acid.

We first created both DNA- and protein-based haplotype networks based on the above definition, collapsing those nodes with identical sequences into one (see Methods). Networks can be made of one or more components. Each component is subgraph in which any two nodes are connected to each other by a path of links. We found that the average size of the largest connected component – commonly referred to as the dominant component – relative to total network size is significantly larger in protein-based networks (12,235 proteins, a fraction 0.975 of the total network) than in DNA networks (15,841 DNA sequences, 0.940 of the total network) (Mann–Whitney U test –

p-value= $7.01 \times 10^{-156}$) (See also Figure 2.2c). Because our statistical analyses focus on the dominant component of each haplotype network and work best if this component comprises as many nodes as possible, we focus on protein-based haplotype networks for the rest of this contribution. The 1,000 genomes dataset we use contains information from 19,744 genes, but we constructed haplotype networks only for those 12,235 protein-coding sequences that showed at least one amino acid variant.

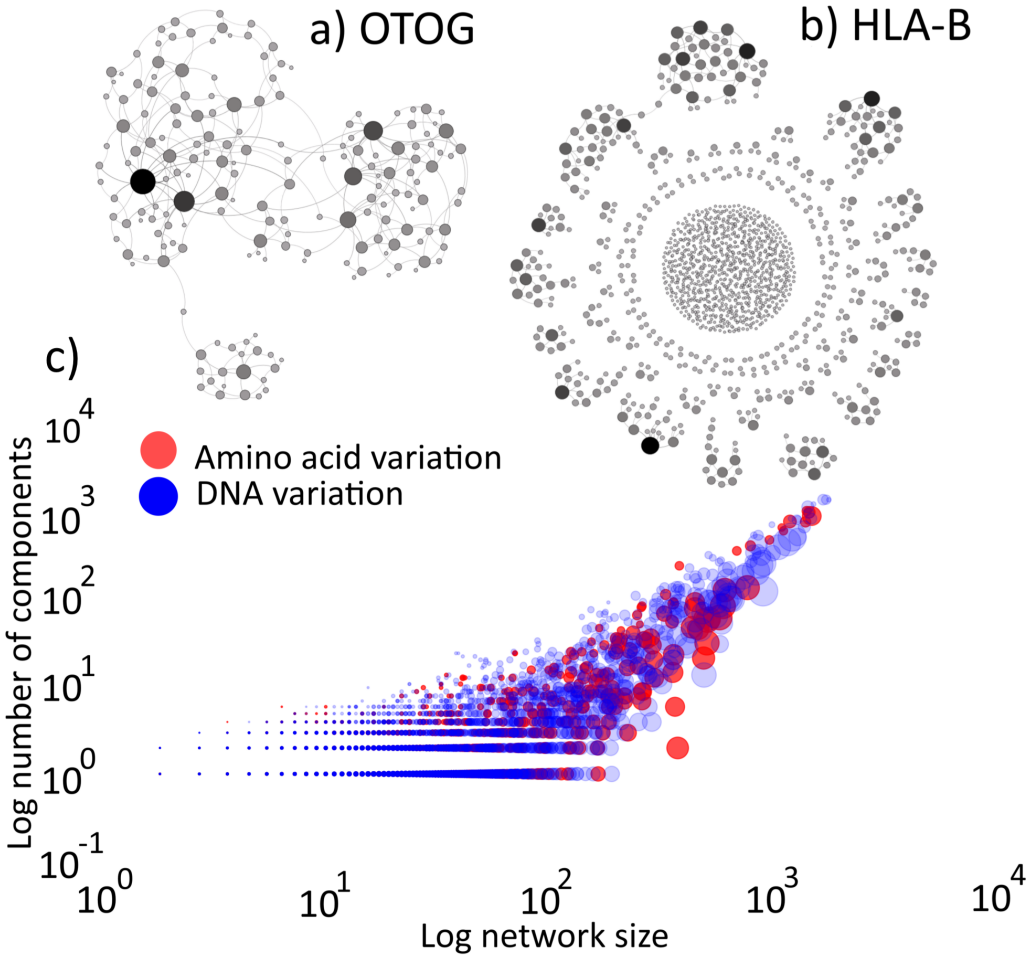


FIGURE 2.2: **Haplotype networks vary greatly in structure among genes. a)** Haplotype network of gene OTOG (Otogelin). Among all protein-based haplotype networks comprising more than 100 sequences, OTOG has the network with the largest dominant component where all nodes fall into this component (181 nodes and a single component). **b)** Haplotype network of gene HLA-B, which is the most fragmented network, with 1,545 nodes in 1,111 components. Circles in a) and b) correspond to different genotypes, while links connect genotypes that differ by a single point mutation. Circle color and size correspond to the degree (number of neighbors) of the node, where darker and larger nodes have a higher degree. **c)** Number of components versus network size for DNA-based (blue circles) and protein-based haplotype networks (red circles). Circle size in c) corresponds to the relative size of the dominant component within each haplotype network.

Figure 2.2 a and b illustrate with two examples that haplotype networks for different genes can show great variation in topology. The left network (Figure 2.2a), derived from the gene *OTOG*, which encodes Otogelin, comprises

181 nodes organized into a single component, whereas the right network, from gene *HLA-B* (Major histocompatibility complex, class I, B) is highly fragmented and has 1,545 nodes in 1,111 different components. (See Figure S2.1 for a different representation of the two networks.)

More generally, Figure 2.2c shows the distributions of the number of components and the size of the largest component. There are 11,155 networks with only a single component, but most of these networks are small, with an average of 5.52 sequences. The network with the most components is the highly fragmented *HLA-B* network with 1,111 components. *HLA-B* is known to be under strong balancing [97] and divergent selection [143], which causes great genotypic diversity. This diversity translates into high network fragmentation, i.e., a network with many components. Some haplotype networks have very large dominant components with up to 552 sequences. However, in most (10,587) networks, the largest component is very small, comprising a maximum of ten sequences. The network with the largest dominant component where all sequences fall into that component is that of *OTOG* (Figure 2.2a).

### 2.2.2 Cycles in haplotype networks

A cycle in a network is a series of links starting from one node and ending with the same node, while passing other nodes along the cycle only once. In haplotype networks constructed from biallelic gene variants, the simplest elementary cycle, i.e. a cycle not decomposable into smaller cycles, is a square. The reason is that cycles with an odd number of links, e.g. triangles or pentagons, are impossible when all SNPs are biallelic. Figure 2.1 shows a square that involves the mutation of a hypothetical DNA sequence at two different sites (positions 10 and 20). Next to each circle (sequence) the nucleotide residues at these positions are indicated, and along the links, the specific nucleotide changes that occurred. If genotype 1 is the most recent common ancestor of its neighbors, then its two neighbors have undergone two different mutations: Specifically, genotype 2 has experienced a change from $A$ to $T$ at position 10 and genotype 3 has a change from $A$ to $G$ at position 20. To form the single genotype 4 from its ancestors, i.e. from either genotype 2 or 3, either genotype 2 has undergone a change from $A$ to $G$ at position 20, or genotype 3 has undergone a change from $A$ to $T$ at position 10, so that the descendants of the two ancestral sequences 2 and 3 become not only more

similar but identical to one another. It is not necessary for both of sequences 2 and 3 to mutate to form genotype 4, but a mutation in either of them can lead to the genotype and form a cycle. Regardless of whether genotype 1 or any other genotype is the common ancestor of the others, a square like this requires convergent sequence change.

In long biopolymer sequences with multiple monomers that evolve through random mutation alone, cycles should be rare, because it is unlikely that mutations become reversed to create sequences more similar to one another. However cycles can be introduced by mutation biases that allow only certain residues to change, or by selection that causes only certain changes to survive, i.e., by evolutionary constraints. Another possibility is recombination, which might occur between genotypes 2 and 3, which can result in genotype 4. The same mechanisms can give rise to longer cycles (e.g., length 6 or 8, Figure S2.2).

Figure 2.3a shows the distribution of the number of squares in all networks. 7,373 of 12,235 networks had no squares. The network with the most squares is that of gene *DNAH11* and contains 1,043 squares. The inset of Figure 2.3a shows the distribution of hexagons and octagons. The networks with the largest number of hexagons (74) and octagons (4) are those of genes *MAP2K3* and *HLA-B*, respectively. Note the small numbers of hexagons and octagons compared to squares. Even though we enumerated elementary cycles up to length eight – beyond that, our analyses become computationally too costly – we focus most of the following analysis on squares, because they are by far the most abundant cycles.

### 2.2.3 Unconstrained or constrained mutation cannot explain the large number of cycles in many networks

Because some amount of homoplasy can occur by chance alone, we wished to determine whether all cycles we observed could have occurred by chance homoplasy. To this end, we created randomized haplotype networks in which the same amount of evolutionary change occurred as in the actual networks.

In our first randomization procedure, we created a (simulated) DNA sequence of the same length as the coding sequence of a gene, and created a haplotype network from it by simulating a pattern of mutations designed to yield a network with the same number of links (number of nonsynonymous

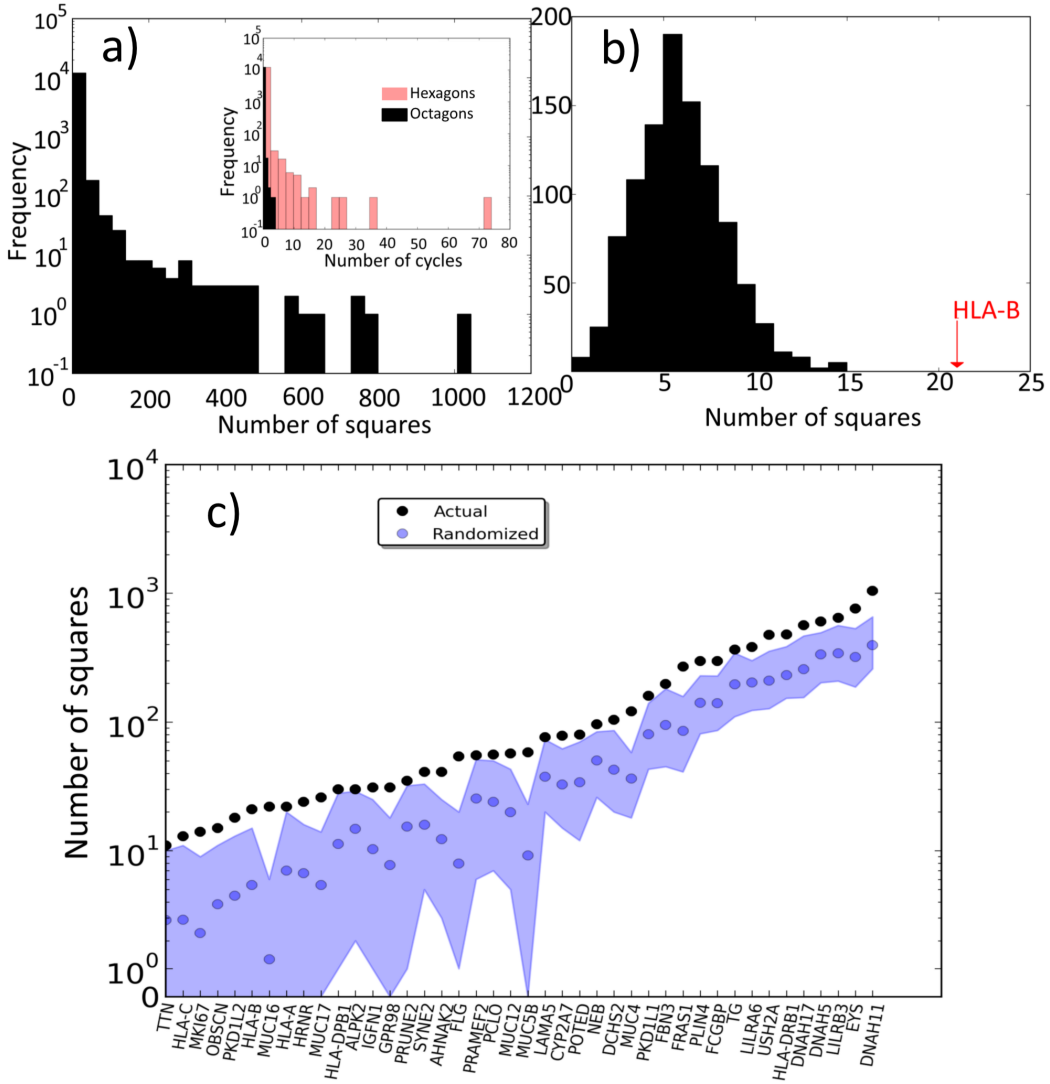


FIGURE 2.3: **Distribution of the number of cycles in all networks and in networks with an excess of squares. a)** Distribution of the number of elementary squares, as well as elementary hexagons and octagons (inset) in all protein-based networks. **b)** Distribution of the number of squares in 1,000 randomized networks derived from the dominant component of the HLA-B network, whose number of squares (21) is indicated by a red arrow. **c)** Number of squares (black circles) in the largest components of the haplotype networks of 42 genes with significantly more squares than expected by chance alone, together with the mean number of squares (blue circles) found in 1,000 randomly generated networks for each gene. Shaded areas depict the maximum and minimum number of squares in the randomized networks. Note the logarithmic scale on the vertical axis.

changes) and the same distribution of degrees (number of neighbors) as the actual network (see Methods). Specifically, we compared the number of cycles in each haplotype network to 1,000 such randomly generated networks, and found 4,267 genes whose actual number of cycles was greater than all of the 1,000 randomly generated networks. Thus, based on this criterion there are 4,267 genes whose total number of cycles cannot be explained by chance homoplasy alone (p-value$\approx 0.001$ – FDR [17] at 0.05) (full list of these genes can be found in the electronic supplementary materials of ref. [211]).

One can argue that this procedure does not take into consideration the actual patterns of variation observed in the data, namely that only a small subsets of sites in any one gene have been subject to mutation, and that all of the sites are biallelic, that is, only two variant nucleotides occur in them. Both patterns arise from the fact that the human population sample is not highly diverged, and that natural selection constrained the evolution of these sequences, i.e., it eliminated some mutations that occurred in them. We thus modified our randomization procedure to reflect these facts (see Methods). With these more conservative criteria, we still found 42 genes (0.34 percent of all genes analyzed) whose haplotype networks have significantly more cycles in their networks than expected by chance alone (Table 2.1). That is, their number of cycles cannot be explained by mutational patterns and purifying selection alone. Figure 2.3b shows, as an example, the number of cycles (21, orange arrow) in the haplotype network of *HLA-B*, which is 6.52 standard deviations greater than the mean number of cycles (5.36) in 1,000 randomized networks (black histogram). Figure 2.3c shows the number of squares in all 42 networks (black circles), together with the mean (blue circles), minimum, and maximum (blue shaded regions) number of squares for 1,000 randomized networks created for each of the 42 haplotype networks.

Figure S2.3 shows the distribution of elementary cycles with length four, six and eight among the 42 genes with an excess of squares, and Figure 2.4 shows the proportion of the sequences that form part of a square in the largest connected component of each gene network. For some genes, such as *POTED* (POTE ankyrin domain family, member D) all sequences form part of a square, and even for genes where the proportion of sequences in a square is low, such as *HLA-C* (major histocompatibility complex, class I, C) and *TTN* (titin), it exceeds 40 percent (Figure 2.4).

We note that the 42 networks with an excess of squares are otherwise very heterogeneous in their properties. They range from the network of *MKI67*

TABLE 2.1: **Genes with an excess of squares in their dominant component.** The number of squares in these genes cannot be explained by random homoplasy or mutational constraints. The middle column cites studies that provide evidence for positive selection, wherever such evidence is available. After FDR correction, the p-value of the statistical test comparing the actual number of cycles against that in 1,000 randomized networks (with random mutations and mutational constraints) is 0.087 for all genes.

| Gene name | Previous evidence of positive selection | Number of squares in the dominant component |
|---|---|---|
| TTN | None | 11 |
| MKI67 | None | 14 |
| OBSCN | None | 15 |
| PKD1L2 | None | 18 |
| MUC16 | None | 22 |
| MUC17 | None | 26 |
| IGFN1 | [209] | 31 |
| GPR98 | [209] | 31 |
| PRUNE2 | None | 35 |
| SYNE2 | None | 41 |
| AHNAK2 | None | 41 |
| HLA-DPB1 | [103–105, 119, 189] | 48 |
| ALPK2 | None | 50 |
| HLA-C | [103–105, 119, 189] | 50 |
| FLG | None | 54 |
| PRAMEF2 | [209] | 55 |
| HRNR | None | 55 |
| MUC5B | None | 58 |
| PCLO | [209] | 67 |
| HLA-A | [103–105, 119, 189] | 67 |
| MUC12 | None | 71 |
| LAMA5 | [209] | 76 |
| CYP2A7 | [47] | 76 |
| HLA-B | [103–105, 119, 189] | 76 |
| POTED | None | 80 |
| NEB | None | 96 |
| MUC4 | None | 121 |
| PKD1L1 | None | 160 |
| FBN3 | [209] | 197 |
| DCHS2 | None | 205 |
| FRAS1 | [209] | 269 |
| PLIN4 | None | 298 |
| EYS | None | 316 |
| FCGBP | [209] | 350 |
| TG | None | 365 |
| USH2A | [209] | 475 |
| LILRB3 | None | 475 |
| LILRA6 | None | 482 |
| DNAH17 | [209] | 494 |
| HLA-DRB1 | [37, 103–105, 119, 189] | 507 |
| DNAH5 | [209] | 602 |
| DNAH11 | None | 1043 |

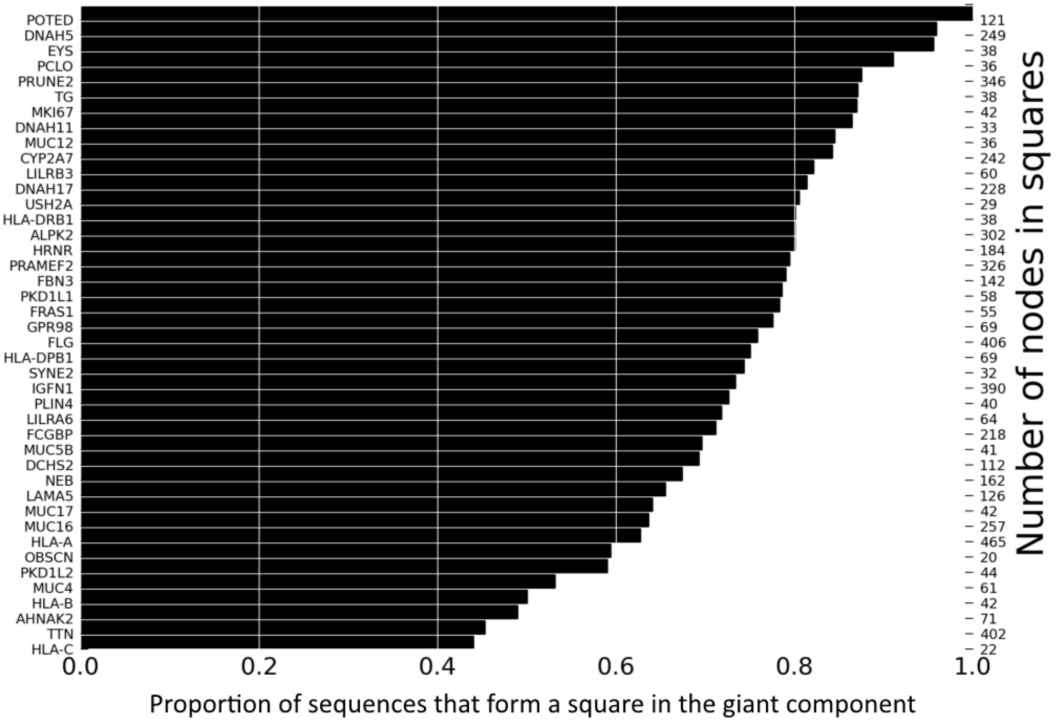


FIGURE 2.4: **Proportion of sequences that are part of a cycle.** Proportion (horizontal axis) and actual number of sequences (right vertical axis) that are part of a square in the giant connected components of haplotype networks for those 42 genes (left horizontal axis) with a significant excess of squares.

(marker of proliferation Ki-67) where only 23 nodes lie in the largest connected component, to the network of *DNAH11* (dynein, axonemal, heavy chain 11), where 538 nodes do (see Figure S2.4 for the distribution of component sizes). Some of the networks have very few components, such as that of *POTED* with a single component, whereas others have many components, such as the highly fragmented *HLA-B* network with 1,111 components (see Figure S2.5 for the distribution of component numbers). Even properties within the largest connected components are heterogeneous. For example, in some networks, such as that of *PKD1L1*, the distribution of the numbers of neighbors of each sequence is highly left-skewed and dominated by sequences with few neighbors, while in others it is more symmetric (*PRAMEF2*, Figure S2.6). Assortativity coefficients, which quantify the tendency of each node to connect to other nodes with a similar number of neighbors, also vary broadly. Some networks are assortative (sequences with many neighbors are adjacent to other sequences with many neighbors), whereas others are disassortative (Figure S2.7).

Gene Ontology (GO) enrichment analysis on biological processes shows several immune system-related processes which are enriched in the 42 genes,

namely "antigen processing and presentation of endogenous peptide antigens" and "interferon-gamma-mediated signaling pathway" (see the electronic supplementary materials of ref. [211] for full results of the analysis and parameters). GO enrichment analysis of molecular functions reveals the two enriched functions "calcium ion binding" and "peptide antigen binding". The "peptide antigen binding" is again associated with the immune system.

Given the strong representation of HLA genes among genes with excess of cycles, we asked how GO enrichment analysis would change if we exclude the HLA genes. We found a single enriched biological process, namely "O-glycan processing", and two enriched molecular functions, namely "calcium ion binding" and "extracellular matrix constituent, lubricant activity".

We also asked whether genes with an excess of cycles preferentially occurred in specific KEGG [115] or Reactome [114] pathways. Six genes were preferentially associated with KEGG pathways. They include TG (thyroglobulin) and the genes in the HLA family. The enriched pathways comprise "Epstein-Barr virus infection", "Autoimmune thyroid disease", "HTLV-I infection", "Viral myocarditis", "Allograft rejection", "Phagosome", "Antigen processing and presentation", "Graft-versus-host disease", "Cell adhesion molecules (CAMs)", "Herpes simplex infection", and "Type I diabetes mellitus".

For Reactome pathways, we found twelve genes enriched in six pathways. The genes include those encoding Mucins, the HLA family and LILR family genes (*MUC4*, *MUC5B*, *MUC12*, *MUC16*, *MUC17*, *HLA-A*, *HLA-B*, *HLA-C*, *HLA-DRB1*, *HLADPB1*, *LILRA6*, and *LILRB3*). The enriched pathways are "Termination of O-glycan biosynthesis", "Interferon gamma signaling", "Endosomal/Vacuolar pathway", "Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell", "Antigen Presentation: Folding, assembly and peptide loading of class I MHC" and "Defective GALNT12 causes colorectal cancer 1 (CRCS1)". We note that both enriched KEGG and Reactome pathways include several immunity-related pathways.

### 2.2.4 Recombination cannot account for an excess of squares in most networks

To exclude the possibility that genetic recombination may account for the excess of squares in some networks, we performed two complementary analyses. First, we simulated for each gene the effect of recombination on haplotype network structure by creating haplotype networks based on a set of sequences that was subject to approximately as many recombination events as occurred in the human population since their common ancestry, as well as to as many mutations as there are links in the network (see Methods). We repeated this process 1,000 times for each gene, creating 1,000 simulated haplotype networks, and counted the number of squares in them. For each of the 42 genes, the empirical network showed more squares than each of the 1,000 simulated networks (Figure S2.8).

In the second analysis, we asked whether gene conversion, a process of unidirectional recombination in which only one of the recombining sequences changes, may have caused the excess of squares [101, 109]. To this end, we used the program GENECONV (version 1.81a) [221] to detect gene conversion among the sequences in the dominant components of the 42 haplotype networks. We used sequences comprising both synonymous and non-synonymous changes to give the program more power in finding gene conversion events. Only one gene showed any sign of gene conversion, and it did so for only two of 114 sequences in *CYP2A7* (cytochrome P450, family 2, subfamily A, polypeptide 7). In sum, based on these analyses, it seems unlikely that recombination can explain the excess of squares we observe in the haplotype networks of 42 genes.

### 2.2.5 Positive selection as a potential cause of squares

Positive selection can be a driver for homoplastic or convergent evolution, where two separate lineages evolve the same trait independently [256]. Because such adaptive homoplasy can occur not only at the phenotypic level [88, 102], but also at the sequence level [32, 199, 278], we wished to find out whether positive selection can help explain the excess of squares we observed in the haplotype networks of 42 genes.

Previous studies had indeed indicated positive selection for at least 17 of the 42 genes [22, 47, 103–105, 119, 189, 209] (Table 2.1). In addition, we used results from a branch-site likelihood test [280] which indicates positive selection based on a ratio $d_N/d_S > 1$ observed along one or more branches of a phylogenetic tree. This test has been applied to vertebrate genes in the Selectome database [209], which indicates that 12 of our 42 genes with abundant squares show patterns of positive selection, either in primates or in the bony vertebrates (Euteleostomi, Table 2.1 and Table S2.2). This number – 12 of 42 – is unlikely to be observed by chance alone ($p = 0.0004$; hypergeometric test, based on 2,125 unique genes in the human genome under positive selection according to Selectome (data provided by the authors of Selectome)). In addition, we used the XP-CLR (cross-population composite likelihood ratio) test for neutrality [37] (see Methods). The test compares different populations to identify rapid changes in a locus' allele frequency that cannot be explained by random drift alone. In applying this test, we used a test statistics [210] precomputed over 2kb sliding windows that covered the human genome, and asked for each of our 42 genes whose haplotype network showed an excess of squares, whether two or more of the windows where the test-statistic indicated the action of positive section ($p = 0.01$) overlapped with the gene (see Methods). By this criterion, six of our 42 genes showed evidence of positive selection in at least one population (Table S2.1 and table S2.3). Overall, 21 of our 42 genes with an excess of squares showed signs of positive selection by at least one of these criteria or by previous work.

We also analyzed patterns of synonymous and non-synonymous changes in more detail. A commonly used indicator of positive selection for two protein-coding DNA sequences is $d_N/d_S$, i.e. the ratio of nonsynonymous changes $d_N$ per nonsynonymous site to synonymous changes $d_S$ per synonymous site. Values of $d_N/d_S > 1$ can indicate positive selection [125, 274]. Unfortunately, $d_N/d_S$ can be computed only for sequences more distantly related than those in our haplotype networks. The reason is that in these networks, adjacent sequence pairs differ only in a single nonsynonymous mutation, and many adjacent pairs do not even show a single synonymous change. More specifically, in the dominant component of our networks, up to 80 percent of sequence pairs do not show a single synonymous mutation (Figure S2.9), and this incidence of synonymous mutation is similarly low in the entire network. Moreover, it has been suggested that for very closely related sequences, $d_N/d_S$ is not a sensitive indicator of positive selection [134]. For

these reasons, we compared the incidence of nonsynonymous and synonymous changes among groups of links (see Methods), reasoning that groups of links with very few synonymous changes might provide hints that some or all members of the group may have been subject to positive selection. Most links show no synonymous changes at all in some networks, which hints that positive selection may have played a role in creating their pattern of diversity (Figure S2.9).

We specifically compared links with no synonymous change inside squares and outside squares. While the difference between the fractions of links without synonymous changes inside squares was not significantly different from those outside squares (Fisher's exact test on $2 \times 2$ contingency tables, Figure S2.10), the average number of synonymous changes on links inside squares was significantly smaller than that outside squares for 14% (6) of the genes (Mann-Whitney U test, p-value$= 0.05$, FDR corrected). Figure 2.5 shows the average number of synonymous changes per edge for links inside squares divided by that for links outside squares. For genes where this ratio is below 1 (red vertical line) the average number of synonymous changes are smaller inside squares than outside squares. Overall, the distribution of synonymous changes among links inside squares and outside squares does not suggest that all incidences of excessive squares are due to positive selection, but it suggests that positive selection may have contributed to this excess for at least some genes.

Using a test based on the hypergeometric distribution [113], we found no significant overlap between the genes that showed evidence of positive selection in the XP-CLR test and those genes among our 42 focal genes that (i) have significantly fewer synonymous mutations inside the squares than outside the squares of their haplotype network (2 common genes) or (ii) had been identified in several previous studies as being subject to positive selection (3 common genes).

### 2.2.6 Balancing selection is not a likely cause of an excess of squares

In a final analysis, we also asked for evidence of balancing selection, which manifests itself as an elevated amount of heterozygosity and can in principle produce squares. Consider, for example, the hypothetical square in Figure

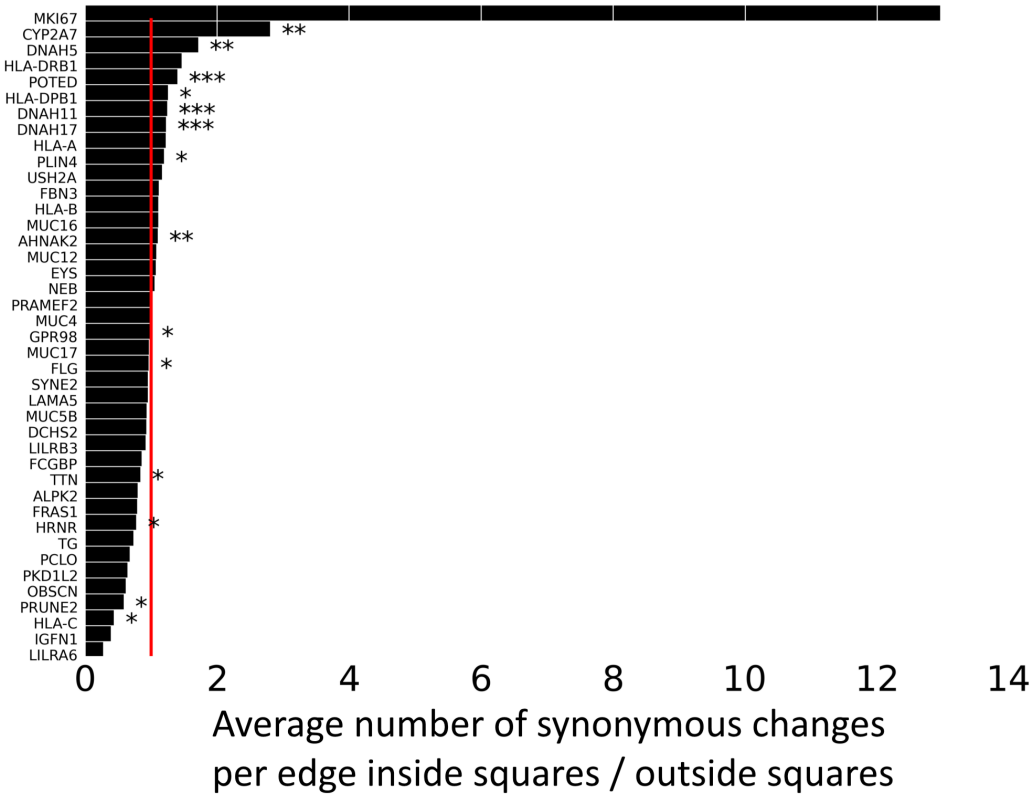


FIGURE 2.5: **Ratio of the average number of synonymous changes per edge for links inside squares relative to links outside squares.** The red line corresponds to a value of this ratio that is equal to one, i.e, links inside and outside squares have the same average number of synonymous changes. Bars that end to the left (right) of this line indicate genes in which the average number of synonymous changes per edge is lower (higher) inside squares than outside squares. *, *, and *** indicate that the difference between the average number of synonymous changes inside versus outside squares is significant at p-values of 0.05, 0.01, and 0.001, respectively (Mann-Whitney U test). The p-values are corrected following [17].

2.6, in which nodes represent hypothetical diploid genotypes. Next to each circle (genotype) the nucleotide residues at positions 10 and 20 are indicated, and along the links, the specific nucleotide changes that occurred for the first of two haplotypes. If genotype 1 is the most recent common ancestor of genotypes 2 and 3, then a substitution at site 20 in the first haplotype of genotype 1 creates genotype 2, and a substitution at site 10 of the first haplotype creates genotype 3. If balancing selection is acting on both sites (10 and 20), individuals 2 and 3 will be favored over individual 1, because they are heterozygous at one of the two sites under balancing selection. A further substitution to genotype 4, would create a double-heterozygous genotype – and a square – that is even more favored by balancing selection.

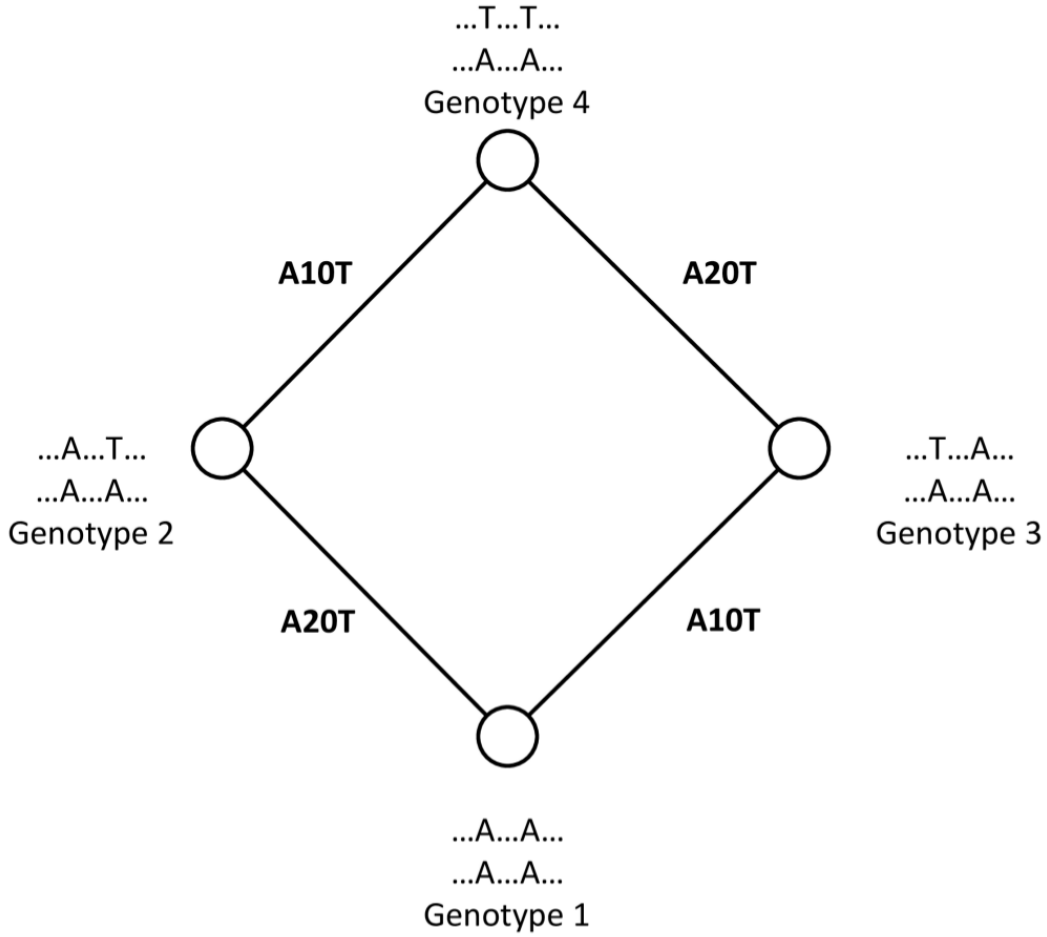


FIGURE 2.6: **Balancing selection can produce cycles.** The example indicates a hypothetical diploid genotype where two nucleotide changes occur at position 10 and 20. Circles (nodes) correspond to genotypes. A link connects two nodes if they differ by a single mutation. Lettering next to each node indicates the nucleotides at which two genotypes differ. Edge labels show changes required to create a genotype from its neighbor, e.g., "A20G" indicates a change from *A* to *G* at position 20 of the first haplotype of the hypothetical genotype. See text for details.

We computed for each gene the fraction of heterozygous individuals averaged over all sites that experienced nonsynonymous changes in at least one

individual of the sample population (see Methods). Among our 42 genes with an excess of squares, we found no significant (Pearson's r, p-value$=0.512$) correlation between the number of squares and heterozygosity. For all 19,744 genes, we found a very small (Pearson's r = 0.066) yet significant correlation ($p = 3.42 \times 10^{-13}$) between heterozygosity and the number of squares in a gene's haplotype network (Figure S2.11). In sum, balancing selection is not a likely explanation for the prevalence of squares in some genes.

### 2.2.7 Multiple genes whose haplotype networks show an excess of squares are implicated in immune functions

Especially prominent among the 42 genes whose haplotype networks show an excess of squares are genes with immune functions. Such genes are also known to be subject to frequent positive selection [185]. For example, five of the 42 genes belong to the human leukocyte antigen (HLA) family. These are the genes *HLA-A*, *HLA-B*, *HLA-C*, *HLA-DPB1*, and *HLA-DRB1*. HLA genes show the highest level of polymorphisms in the human genome [119, 217], and display multiple signatures of positive selection, including a high $d_N/d_S$ in antigen-recognition sites, trans-species polymorphisms, high levels of heterozygosity, as well as long range haplotypes, a key signature of recent positive selection [119].

Five more among the 42 genes with an excess of squares encode mucins, which are important for the immune response, because they help form mucus that can prevent pathogen entry, and cooperate with antibodies to fight pathogens [23, 69, 100]. These are *MUC4*, *MUC5B*, *MUC12*, *MUC16* and *MUC17*.

Two more among the 42 genes, *LILRB3* and *LILRA6*, encode leukocyte immunoglobulin-like receptors, which cooperate with MHC proteins. *LILRB1*, another member of this family, has co-evolved with HLA, which is under positive selection in sub-Saharan population [186]. Another immune-relevant gene among the 42 genes is *FCGBP*, which may play an important role in immune protection and inflammation in the intestines of primates [94].

## 2.3 Discussion

We show that the haplotype networks of 42 genes display a significant excess of squares that cannot be explained by chance homoplasy, genetic recombination, or balancing selection alone. This leaves constrained evolution as a prominent candidate cause, which limits the diversity of alleles that are generated or preserved in a sequence. While such constrained evolution can have multiple causes [226], strong purifying and positive selection are most relevant for the kind of data we analyze.

Strong purifying selection may play a role in the occurrence of squares, because we observed significantly fewer squares for many genes in our randomization tests when we allowed the whole protein coding sequence to change, and when we permitted substitutions to any nucleotide. In addition, some of the genes with an excess of squares may have experienced positive selection. First, up to 80% of links in the dominant component of some of these genes do not have any synonymous mutations at all (Figure S2.9). Furthermore, six of the genes with an excess of squares (14%) have significantly more synonymous changes outside their squares than inside them (Figure 2.5). In addition, six genes contained at least two adjacent windows with a significantly high value of the XP-CLR test statistic that can indicate positive selection (Table S2.1). Moreover, previous studies have suggested that 17 of the 42 genes with an excess of squares have been subject to positive selection (Table 2.1). Finally, multiple genes among those with an excess of squares are involved in immune functions, which are frequently subject to positive selection ([185]).

In addition, it is relevant that there is a mounting number of known genes where convergence at the sequence level has been caused by positive selection. For example, sequence convergence occurred in the peptide-binding regions of human and mouse class Ib genes in the major histocompatibility complex (MHC), the same gene family in which five members show an excess of squares in our study [276]. The motor protein Prestin, which is involved in the mammalian auditory system, has experienced adaptive sequence convergence between echolocating bats and echolocating dolphins [147]. Two other genes involved in the mammalian auditory system, *Tmc1* and *Pjvk*, also have experienced convergence due to positive selection [52]. In addition, whole genome sequencing of four bat species showed extensive genome-wide convergence among these taxa [199]. Moreover, extensive convergent evolution

occurred between snake and agamid lizard mitochondrial genomes, much of which may be adaptive [32].

Our analysis is based on some 1,000 human genomes, which raises the question how its results might be affected as the size of the available data set increases. Most importantly, a larger data set would lead to larger and more connected networks. Our analysis is focused on the largest connected component of each network, and increasing the size of the largest connected component could lead to more cycles just by chance alone. Indeed, larger connected components of a haplotype network in our data set also contain more cycles (Figure S2.12). This pattern also extends to those networks with a significant excess of cycles. Specifically, dominant component sizes are significantly larger for networks that have a significant excess of cycles than for the remainder of the haplotype networks (Figure S2.13). Conversely, a higher fraction of genes with an excess of cycles have large dominant components ($> 100$ nodes). These observations suggest that increasing the size of our data set might not just increase the overall number of cycles, but also the number of haplotype networks with an excess of cycles. In other words, it would increase the sensitivity of our analysis.

A recent study [26] has shown that HLA genes show reference allele bias in the 1,000 genomes data. Removing these alleles from the dataset could in principle lead to smaller dominant components in the HLA networks and hence to fewer cycles. However, this is unlikely to materially affect our observations, because the largest components, with one exception, comprise a small fraction of the HLA networks (0.05, 0.26, 0.09, 0.60 and 0.04 for *HLA-B*, *HLA-DPB1*, *HLA-A*, *HLA-DRB1* and *HLA-C*, respectively). Thus, most removed alleles would fall into other components, and their removal would thus not affect our giant-component-based analysis.

We used two different approaches for analyzing the effect of recombination on the occurrence of cycles in gene networks. We used a test of gene convergence on actual genetic variation data, and we also constructed random networks with mutation and recombination. We did not find any strong evidence of gene convergence in any of the 42 genes with an excess of cycles. Random networks with recombination did not produce as many cycles as we observed in our gene networks. We should note that in the recombination process we implemented, each sequence recombines with a maximum of one other sequence. This might have an effect on the number of cycles occuring in a random network compared to a recombination procedure where

recombination can occur among multiple sequences. It would be interesting to explore other recombination procedures too.

In sum, while we have not been able to explain the abundance of squares conclusively, we suggest that a mix of constrained evolution through purifying selection and positive selection may be responsible. As data from more and more individuals from the global human population become available, it will be possible to disentangle these causes. Such data may also help explain the great differences in haplotype network structure among the human genes we characterized here.

## 2.4 Conclusion

We explored a novel way of representing human genetic variation data through a network-based approach whose strengths are complementary to phylogenetic trees. Despite the fact that the genes in the genomes we analyze have a shared phylogenetic history, they show very diverse properties in their haplotype networks. Specifically, these networks show different numbers of genotypes (Figure 2.2c), different extents of fragmentation (Figure S2.5), different degree distributions (Figure S2.6), and different assortativity (Figure S2.7). Our analysis focuses on the feature of these networks that cannot be easily represented in phylogenetic trees, i.e., cycles. Phylogenetic trees are acyclic, and thus not ideally suited to represent evolutionary histories more complex than direct descent, such as allopolyploidization, convergent evolution, sexual reproduction, recombination and horizontal gene transfer. Such events can transform a tree-like evolutionary history into a reticulate network. Haplotype networks can represent such reticulation, and can thus complement phylogenetic trees in their ability to represent and describe evolutionary processes.

## 2.5 Methods

### 2.5.1 Construction of haplotype networks

We focused our analysis on haplotype networks built from amino acid changing (non-synonymous) mutations of all genes in the human genome, and

supplemented this analysis with data on synonymous mutations. The data we use consists of SNPs called from sequencing of 1,092 individuals by the 1,000 genomes project phase I [168]. First we downloaded variant call format (VCF) files [51] containing all genotypic variants for all 1,092 individuals, as well as the functional annotation of the variants (build 23.11.2010) provided by the 1,000 genomes project. At this stage we had 22 VCF files, one for each of the 22 autosomal chromosomes.

Next, using the software VCFtools [51], we filtered the VCF files by removing all sites with a "FILTER" tag other than "PASS", as well as indels, non-phased variants, and all variants with a minor allele frequency smaller than 0.01. Analyzing VCF files after filtering, we found no SNP with more than two alleles, which is why all our analyses are based on biallelic SNPs. Subsequently, we used the previously obtained functional annotation information to create three VCF files for each gene, which contained nonsynonymous, synonymous, and both synonymous and non-synonymous SNPs in the gene's protein coding region.

The networks we analyze are built on the basis of haplotypes, i.e., we considered for each individual its two haploid genotypes separately. Each network is a graph whose nodes are haplotypes, and two haplotypes are connected by links if they differ in a single SNP. Overall, we analyzed 2,184 haplotypes, and established a separate haplotype network for each of 17,744 human genes. We constructed and analyzed all networks with the help of the iGraph package for Python (version 0.6.5) [48], and visualized them using Gephi (version 0.8.2-beta) [16].

For our analysis of protein-based haplotype networks, we merged two haploid genotypes into a single node of the network if they had identical haplotypes based on their non-synonymous SNPs. Some of our analyses required us to compute the number of synonymous changes between adjacent nodes of these networks, and because a node does not necessarily correspond to a unique haplotype, this number is also not unique – different haplotypes encode the same protein but they may differ at synonymous sites. Wherever this was the case, we used in our analysis the average number of synonymous changes along a link, computed by enumerating synonymous changes between all possible pairs of haplotypes for the incident nodes.

### 2.5.2 Analysis of cycles and other network properties

Cycles in a haplotype network are paths that start and end at the same node, while visiting every other node in the path exactly once. We note that in a haplotype network of biallelic SNPs, no cycles of uneven length are possible. We first focused on cycles of length four, i.e., squares, and calculated their number through exhaustive enumeration. Specifically, we started from any one node and walked from there to all its neighbors, the neighbor's neighbors, and so on, avoiding previously visited nodes, until we had visited five nodes. Any sequence of five nodes is a square if the first and last nodes in the sequence are identical. Repeating the same procedure from all nodes in the network allowed us to enumerate all squares (not double-counting squares that we had found more than once). We applied the same approach to find longer cycles of length six and eight. We call such a longer cycle elementary, if it is not decomposable into shorter cycles, and we verified this property for each longer cycle.

### 2.5.3 Randomized haplotype networks

To ask whether the number of cycles in an empirically observed haplotype network is greater than expected by chance alone, we created randomized haplotype networks for each gene. More specifically, this analysis focused on the largest component of each gene's haplotype network, which comprises on average 97.5 percent of a network's nodes.

A randomized network may have fewer or more cycles than the actual network. Consider the hypothetical square $uvyw$ in a haplotype network, where $v$ and $w$ are located at two diagonally opposed corners of the square. In creating a random network, we might start from a node (sequence) $u$, mutate the sequence twice at random to create nodes $w$ and $v$, and then mutate $w$ and $v$ once more (into $w\prime$ and $v\prime$), so that we have created a random network of four links. If $w\prime$ and $v\prime$ are not identical to each other and to the sequence $y$ in the square this random network is not cyclic, whereas the actual four-node network is. (The opposite is also possible, where the randomization process creates a cycle where the actual network does not contain one.)

We performed two types of randomization analyses, one only with mutation and the other with mutation and recombination. Before we explain these

analyses, we highlight a methodological detail. As we mentioned in the introduction of the paper, three substitutions are necessary to observe a square (and four are possible). In our randomization analyses described below, we always use four mutations, which is a statistically conservative choice. It allows the links in randomized networks that have no corresponding edge in the data-based networks, and some of these links can lead to the creation of additional cycles. Thus, the number of cycles expected by chance alone (i.e., in randomized networks) will be somewhat higher with our procedure than in a population evolving subject to the assumptions we make below. This renders any assertion that a haplotype network contains more cycles than expected by chance statistically conservative.

**Randomization with mutation**

In a first randomization analysis, we aimed to create, for each gene, networks with the same number of nucleotide changes as the gene's actual network. To construct such a random network, we began with a single random sequence that we then mutated iteratively. Specifically, we chose a random node $u$ from the actual network and assigned a random sequence to it. Then we mutated the sequence as many times as $u$ had neighbors in the actual network, and assigned each mutated sequence to one of the neighbors. Next, we cycled over each of these neighbors, and for each such neighbor $v$ we mutated its assigned sequence as many times as the number of neighbors $v$ had in the actual network. We repeated this simulated mutation process until all nodes in the original network had been visited, and for as many mutations as there were links in the original network, thus creating a random network based on the same number of links as the original network. Overall, for each gene we created 1,000 such random networks, and counted the squares in all of them.

In this process, we used two different kinds of starting sequences. The first was a random DNA sequence with the same length as the full length protein coding DNA sequence, where each of the four nucleotides was equally likely to occur at every site. Because most human genes have multiple transcripts and the transcripts may overlap with each other, we considered the total length of a gene's protein coding DNA as the stretch of DNA that was covered by at least one transcript. We allowed every site to mutate into one of the three other nucleotides, as long as the mutation was nonsynonymous. To

create nonsynonymous mutations, we chose a transcript for the gene at random, and mutated a random nucleotide site within that transcript. We mutated this nucleotide to some other randomly chosen nucleotide, and determined whether the change was nonsynonymous. If so, we kept the mutation, otherwise we repeated this procedure until we had found a nonsynonyomus change.

The second kind of starting sequence takes into account the observed pattern of variation in the sequences under consideration. This sequence comprised only as many nucleotide monomers as there were sites with nonsynonymous changes in a gene's protein coding amino acid sequence. Moreover, since our data comprises only biallelic SNPs, we allowed each site in this sequence to convert only between two types of residues. We note that relaxing either assumption would lead to even fewer squares in a randomized network than we found. Thus, a randomization test based on this starting sequence is highly conservative.

Since more than 1,000 randomization tests for each network were not computationally feasible, the p-values of our tests could not be smaller than 0.001. To correct for multiple testing, we first assigned a p-value of 0.001 to those networks that had more squares than each of their corresponding 1,000 randomized networks. Then we adjusted p-values of all the networks that had at least one square (4,862 networks) using the procedure of Benjamini and Hochberg [17]. When building networks from full-length protein coding sequences, and from shorter sequences that reflect only the number of polymorphic sites, the adjusted p-values of genes whose randomized networks never had as many or more squares than the actual network were $p = 0.001$ and $p = 0.087$, respectively.

**Randomization with recombination**

To assess whether recombination can help explain the number of squares in human haplotype networks, we constructed, for each gene, 1,000 randomly generated networks that incorporate recombination during their construction, and determined the distributions of squares in these networks. To build a random network with recombination, we started with a collection or "population" of diploid sequences, whose size was half of the number of sequences in the dominant component of the focal gene's haplotype network. (We chose this size because we conceive of these sequence pairs as diploid

"individuals" from which we would later construct a random haplotype network.) All individuals started with the same homozygous randomly generated sequence pair, which was as long as the number of nonsynonymous polymorphic sites in the gene. For each such sequence pair, we determined a number of mutation and recombination events that they were to undergo, as described further below. We then mutated each individual and recombined the two copies of its genome as many times as specified by these numbers. Subsequently, we randomly paired individuals and created each of two "offspring" from each pair by randomly sampling (with replacement) a haplotype from each parent in the pair to an offspring. We used these offspring to construct the random haplotype network, connecting two haplotypes if they differed by a single nonsynonymous mutation.

In this procedure, we wanted to generate a total number of mutations (for all sequences in the population) that was equal to the number of links (nonsynonymous changes) in the dominant component of the focal haplotype network. To this end, we first determined the average number of mutations per individual $M$ as the total number of desired mutations divided by the number of haplotypes in the population. If $M$ was an integer, we mutated each individual exactly $M$ times. If $M$ was a decimal number and $M < 1$, then we introduced a single mutation into the individual with probability $M$, and no mutation with probability $1 - M$. If $M$ was a decimal number and $M > 1$, then $M$ lay in the interval $(k, k + 1)$, where $k$ is some integer. In this case, we introduced $k + 1$ mutations into the individual with probability $M - k$, and $k$ mutations with probability $1 - (M - k)$. We introduced each mutation into each haplotype by choosing a random site from the sequence and changing its nucleotide. To keep the variational constraints imposed by biallelic variation at each site, we only allowed each nucleotide to mutate to one other nucleotide.

If two sequences were to be recombined in the simulation, then recombination took place after mutation, and occurred between haplotypes of each sequence pair. To recombine a sequence $v$ with a sequence $w$, we chose a random position in the sequence, and then replaced all the sites after that position in sequence $v$ with residues in sequence $w$, and also replaced all sites after that position in sequence $w$ with residues in sequence $v$. If two sequences were to be recombined more than once (see below), we repeated this process.

We next describe how we determined the number of recombination events

for each haplotype network, where we aimed at introducing as many recombination events as are likely to have taken place in a gene, based on available polymorphism data. We calculated the fraction r of sequence pairs to be recombined once for each gene and used it for all random networks to be created for that gene. To obtain r, we first multiplied the average per-generation recombination rate in the human genome (0.952 $cM/Mb$ per generation, calculated based on data from [129]) with the number of generations since the sequences in our data set may have shared a common ancestor. To estimate this number of generations, we used the number of synonymous mutations observed in each gene in our data set. Specifically, we used the following relationship

$$generations\ to\ common\ ancestory = \frac{s}{L \times \mu \times N_e} \tag{2.1}$$

where $S$ in the numerator designates the observed number of synonymous sites for that gene (determined using the filtered VCF files from the 1,000 genomes data). In the denominator, $L$ is the length of the gene, including introns, as retrieved from Biomart (version 0.7, [118]), $\mu$ is the average human mutation rate per nucleotide ($1.1 \times 10^{-8}$) [61], and $N_e$ is the effective population size, for which we used a value of $N_e = 10,000$ [238].

After having computed the estimated number of recombination events for each gene, we divided this number by the sample size of our data (1,092) to obtain the number of recombination events $r$ per sequence pair. If $r$ was an integer, then each sequence pair would undergo exactly $r$ crossing over events. If $r$ was a decimal number and $r < 1$, then we introduced a single crossing over event into the pair with probability $r$, and no such event with probability $1 - r$. If $r$ was a decimal number and $r > 1$, then $r$ lay in the interval $(k, k + 1)$, where $k$ is some integer. In this case, we introduced $k + 1$ crossing over events with probability $r - k$, and $k$ crossing over events with probability $1-(r-k)$. Overall, our recombination procedure ensures that the number of recombination events is approximately the same as expected for a set of sequences with comparable diversity as that observed in our data.

In addition to the parameters described above, we constructed randomized network with higher recombination rates, to account for heterogeneous recombination rates across the genome, or higher effective population size, to account for higher effective population size for some genes such as HLA genes. Specifically, we constructed randomized networks with a 10 times

higher effective population size, i.e. 100,000 individuals, and randomized networks with twice the recombination rate that was used initially in the paper. The new recombination rate is 1.90 $cM/Mb$.

The changes in recombination rate and effective population size have not changed the final results. All the genes that were tested had more cycles in the dominant component of their networks than any of the 1,000 randomized networks. Figure S2.14 shows the mean and range of cycle count in the new randomized networks compared with the cycle count in original networks of the genes.

If many synonymous mutations are shared among sequences, the procedure from equation 2.1 would overestimate the number of needed recombination events if we simply counted the number S of synonymous changes across links of a haplotype network. To find out whether this could be the case, we computed the number of synonymous changes that are shared among links. (We note that each node in a haplotype network can correspond to multiple sequences that encode the same amino acid sequence, but may differ in synonymous changes, such that each edge can have multiple sets of associated synonymous changes.) To this end, we counted the fraction of synonymous changes on each edge that are also present in some other edge of the network. This fraction is small, with a median of 0.0459 and a mean of 0.0613. Thus, shared ancestry of synonymous changes is unlikely to confound our estimation of the number of recombination events.

### 2.5.4 XP-CLR neutrality test

We chose to use the XP-CLR (cross-population composite likelihood ratio) test [37] to test for neutral sequence evolution, because this test is robust to demographic history and recombination rate heterogeneity, and it detects both recent and ancient selective sweeps [37]. Briefly, the test searches for regions in the genome in which allele frequencies have changed too quickly to be explained by genetic drift. We used test statistics calculated for 2 kbp sliding windows calculated by [210] for the whole genome, based on the 1,000 genomes data [168]. Specifically, we performed this test for three populations, namely the CEU population (Utah Residents with Northern and Western European ancestry), the CHB population (Han Chinese in Beijing, China) and the YRI population (Yoruba in Ibadan, Nigeria) [61], which amounts to

six possible population pairs and thus six calculations of the test statistics. To find the significance of the test statistics for any one gene of interest, we rank-ordered all the 2kb windows in the genome by p-value, omitting windows with a value of the statistic equal to zero, i.e., lacking information. To identify candidate genes subject to positive selection, we determined which windows overlapped with each one of the 19,221 human genes. Only about three percent of the windows that overlapped genes had a value of the statistic that indicated positive selection (at $p = 0.05$), but these windows overlapped with nearly 20% of genes. This suggests that using this criterion to identify genes subject to positive selection would lead to a high false-discovery rate of positively selected genes. Therefore, we chose a more conservative criterion of calling only those genes subject to positive selection where at least two contiguous windows showed a significantly high test statistic ($p = 0.01$). According to this criterion, only two percent of genes were subject to positive selection in each of the six population pairs.

### 2.5.5 Calculating heterozygosity

To calculate the heterozygosity of any one gene, we used not haplotypes but (diploid) genotypes, and calculated the fraction of heterozygote individuals in our data set at each site where a non-synonymous amino acid change had occurred. We used the average of this value over all sites as our measure of the gene's heterozygosity.

### 2.5.6 Gene enrichment analysis

We used the g:Profiler web tool (Version: r1622_e84_eg31) [214] to ask if any gene ontology (GO) categories of biological processes and molecular functions or any pathways are significantly enriched in the 42 genes with a significant excess of squares in their haplotype network. In this analysis, we used default parameters of the tool, with two exceptions. First, we only searched for enrichment among GO biological processes and molecular functions, as well as among KEGG and Reactome pathways. Second, we set the hierarchical filtering of results, which provides a compact data representation, to "best per parent (moderate)". GO terms are hierarchically related, and not filtering them hierarchically leads to unmanageably long and indiscriminate lists

of enriched functions. The filtering uses the parent-wise grouping of significant terms and results in shorter GO output that is easier to analyze. Details of test results and parameters can be found in the electronic supplementary material of ref. [211].

### 2.5.7 Gene conversion analysis

We used the GENECONV software on Linux (version 1.81a) [221] to detect gene conversion (with default parameters). The sequences that we supplied to the program included the haplotypes that comprised the dominant component of the gene and included both synonymous and nonsynonymous changes.

## 2.6 List of abbreviations

**XP-CLR:** Cross-population composite likelihood ratio test
**DNAH5 (ENSG00000039139):** dynein, axonemal, heavy chain 5
**USH2A (ENSG00000042781):** Usher syndrome 2A
**TG (ENSG00000042832):** thyroglobulin
**SYNE2 (ENSG00000054654):** spectrin repeat containing, nuclear envelope 2
**DNAH11 (ENSG00000105877):** dynein, axonemal, heavy chain 11
**PRUNE2 (ENSG00000106772):** prune homolog 2 (Drosophila)
**MUC5B (ENSG00000117983):** mucin 5B, oligomeric mucus/gel-forming
**PRAMEF2 (ENSG00000120952):** PRAME family member 2
**LAMA5 (ENSG00000130702):** laminin, alpha 5
**FRAS1 (ENSG00000138759):** Fraser extracellular matrix complex subunit 1
**FBN3 (ENSG00000142449):** fibrillin 3
**FLG (ENSG00000143631):** filaggrin
**MUC4 (ENSG00000145113):** mucin 4, cell surface associated
**MKI67 (ENSG00000148773):** marker of proliferation Ki-67
**OBSCN (ENSG00000154358):** obscurin, cytoskeletal calmodulin and titin-interacting RhoGEF
**TTN (ENSG00000155657):** itin
**PKD1L1 (ENSG00000158683):** polycystic kidney disease 1 like 1
**IGFN1 (ENSG00000163395):** immunoglobulin-like and fibronectin type III domain containing 1

**ADGRV1 (ENSG00000164199):** adhesion G protein-coupled receptor V1
**POTED (ENSG00000166351):** POTE ankyrin domain family, member D
**PKD1L2 (ENSG00000166473):** polycystic kidney disease 1-like 2 (gene/pseudogene)
**PLIN4 (ENSG00000167676):** perilipin 4
**MUC17 (ENSG00000169876):** mucin 17, cell surface associated
**MUC16 (ENSG00000181143):** mucin 16, cell surface associated
**NEB (ENSG00000183091):** nebulin
**AHNAK2 (ENSG00000185567):** AHNAK nucleoprotein 2
**PCLO (ENSG00000186472):** piccolo presynaptic cytomatrix protein
**DNAH17 (ENSG00000187775):** dynein, axonemal, heavy chain 17
**EYS (ENSG00000188107):** eyes shut homolog (Drosophila)
**HLA-DRB1 (ENSG00000196126):** major histocompatibility complex, class II, DR beta 1
**DCHS2 (ENSG00000197410):** dachsous cadherin-related 2
**HRNR (ENSG00000197915):** hornerin
**CYP2A7 (ENSG00000198077):** cytochrome P450, family 2, subfamily A, polypeptide 7
**ALPK2 (ENSG00000198796):** alpha-kinase 2
**HLA-C (ENSG00000204525):** major histocompatibility complex, class I, C
**LILRB3 (ENSG00000204577):** leukocyte immunoglobulin-like receptor, subfamily B (with TM and ITIM domains), member 3
**MUC12 (ENSG00000205277):** mucin 12, cell surface associated
**HLA-A (ENSG00000206503):** major histocompatibility complex, class I, A
**HLA-DPB1 (ENSG00000223865):** major histocompatibility complex, class II, DP beta 1
**HLA-B (ENSG00000234745):** major histocompatibility complex, class I, B
**LILRA6 (ENSG00000244482):** leukocyte immunoglobulin-like receptor, subfamily A (with TM domain), member 6
**FCGBP (ENSG00000090920):** Fc fragment of IgG binding protein

## 2.7 Supplementary figures

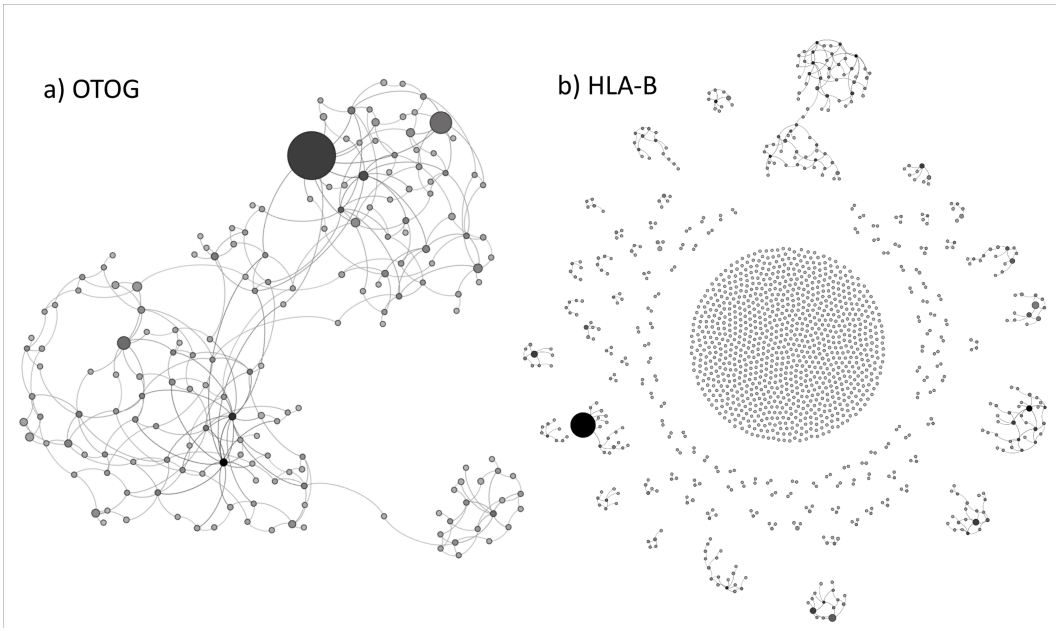


FIGURE S2.1: **Illustration of two haplotype networks, one highly connected and the other highly fragmented. a)** Haplotype network of gene OTOG (Otogelin). Among all protein-based haplotype networks comprising more than 100 sequences, OTOG has the network with the largest dominant component where all nodes fall into this component (181 nodes and a single component). **b)** Haplotype network of gene HLA-B, which is the most fragmented network, with 1,545 nodes in 1,111 components. Circles in a) and b) correspond to different genotypes, while links connect genotypes that differ by a single point mutation. Circle color corresponds to the degree (number of neighbors) of the node, where darker nodes have a higher degree, and circle size corresponds to the number of haploid individuals with that genotype, where larger nodes are shared among more haploid individuals.

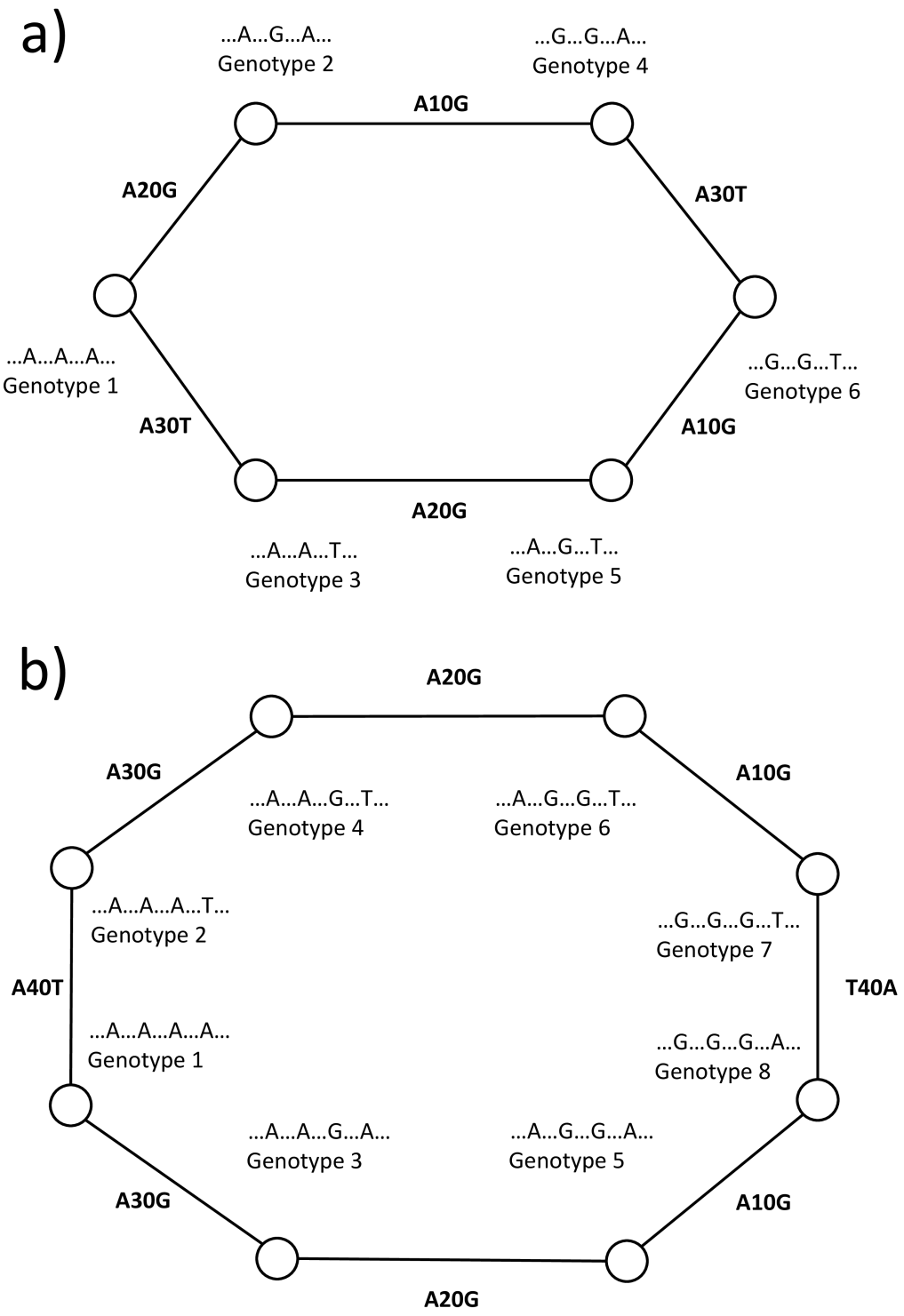


FIGURE S2.2: **Cycles in haplotype networks illustrated with the example of a hexagon and an octagon.** Circles (nodes) correspond to genotypes. A link connects two nodes if they differ by a single mutation. Lettering next to each node indicates the nucleotides at which two genotypes differ. Edge labels show changes required to create a genotype from its neighbor, e.g., "A20G" indicates a change from A to G at position 20 of the hypothetical sequence. **a)** hypothetical hexagon in which six nucleotide changes occur, two each at positions 10, 20 and 30. If one starts from genotype 1, this genotype mutates twice and produces genotypes 2 and 3. Those genotypes in turn mutate to produce genotypes 4 and 5. Then either genotype 4 mutates at position 30 from A to T, or genotype 5 mutates at position 10 from A to G, or both of these mutations happen together, to produce genotype 6. This can be happen when there are evolutionary constraints that restrict other mutations. Recombination can also be responsible for this pattern. This pattern will be the same if one starts from any other node. **b)** hypothetical octagon in which eight nucleotide changes occur, two each at positions 10, 20, 30, and 40. Same pattern that was explained for a) can be explained here, with the only difference that there are more positions that are mutating.

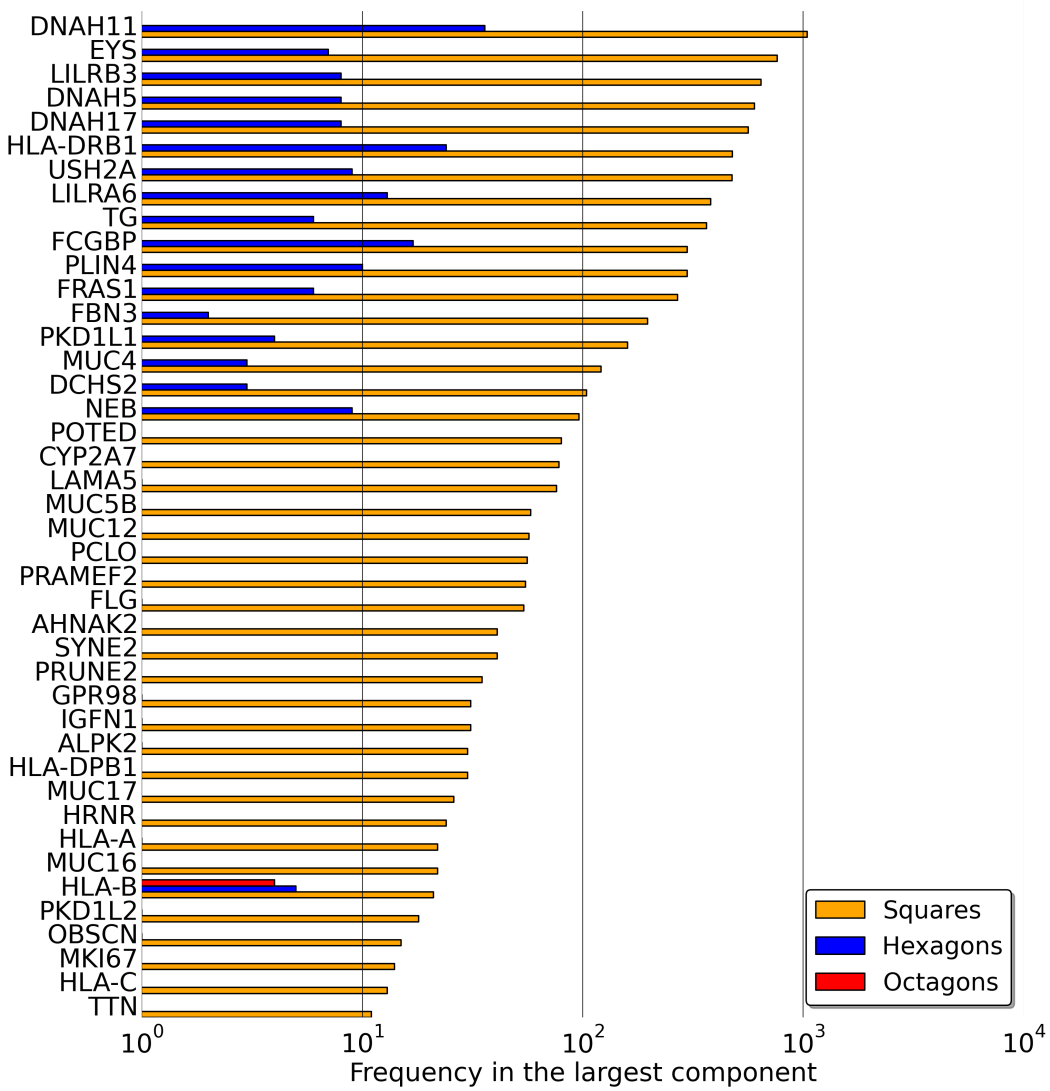


FIGURE S2.3: **Frequency of squares, hexagons and octagons among the 42 genes with an excess of cycles.** The plot shows the frequency of elementary cycles of length 4, 6 and 8 in the dominant component of genes with an excess of squares in their haplotype network. Note that the apparent discrepancy to Figure 2.3a comes from the fact that Figure 2.3a shows cycle numbers for haplotype networks of all genes.

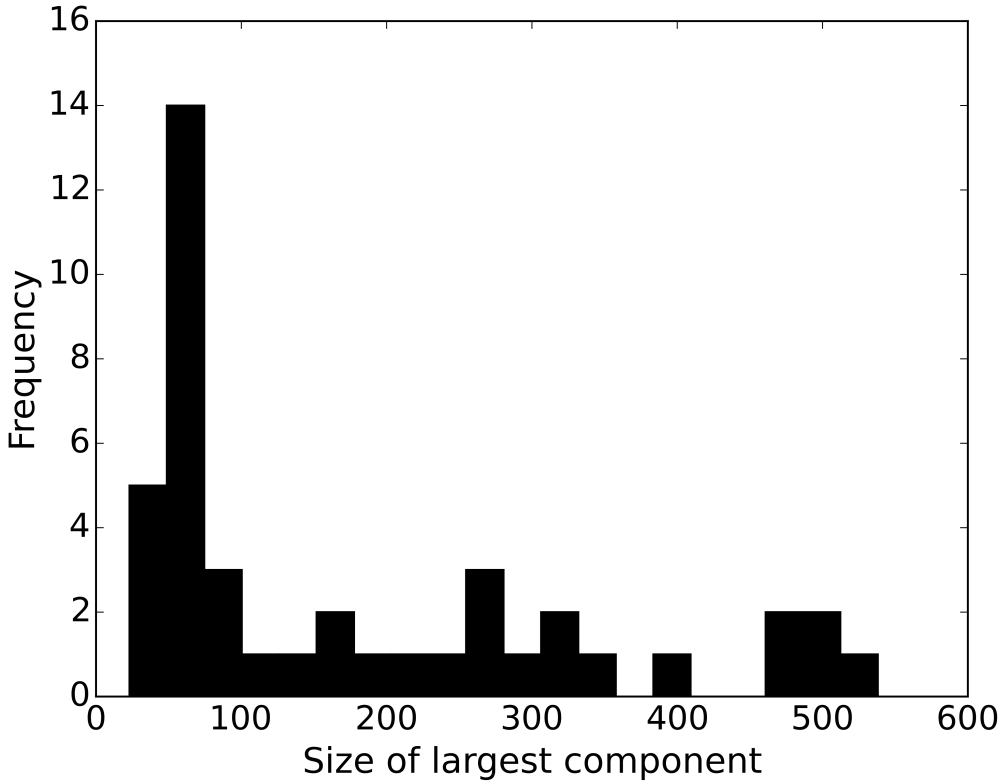


FIGURE S2.4: **Distribution of the size of the largest component in haplotype networks of 42 genes with an excess of squares in the largest component.** The smallest dominant component occurs in the network of *MKI67* (marker of proliferation Ki-67) with only 23 nodes, and the largest one occurs in the network of *DNAH11* (dynein, axonemal, heavy chain 11) with 538 nodes.

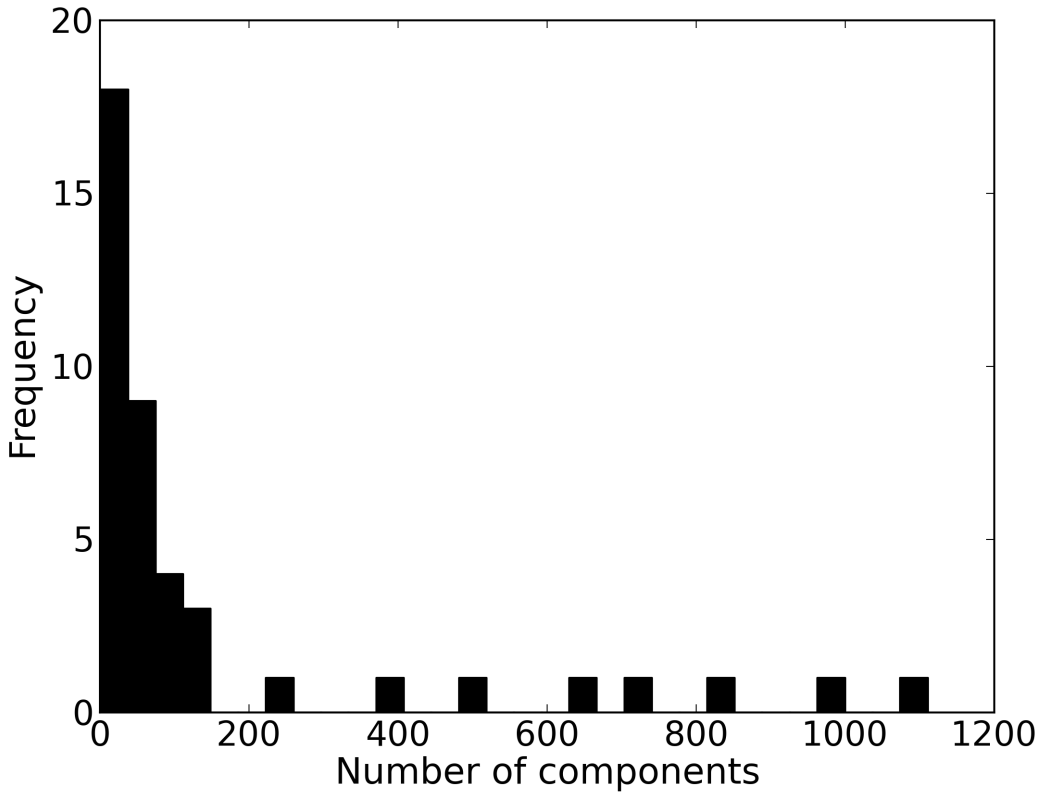


FIGURE S2.5: **Distribution of the number of components in haplotype networks of 42 genes with an excess of squares in their largest component.** The number of components ranges from one for gene *POTED* (POTE ankyrin domain family, member D) to 1,111 for the highly fragmented network of *HLA-B*.

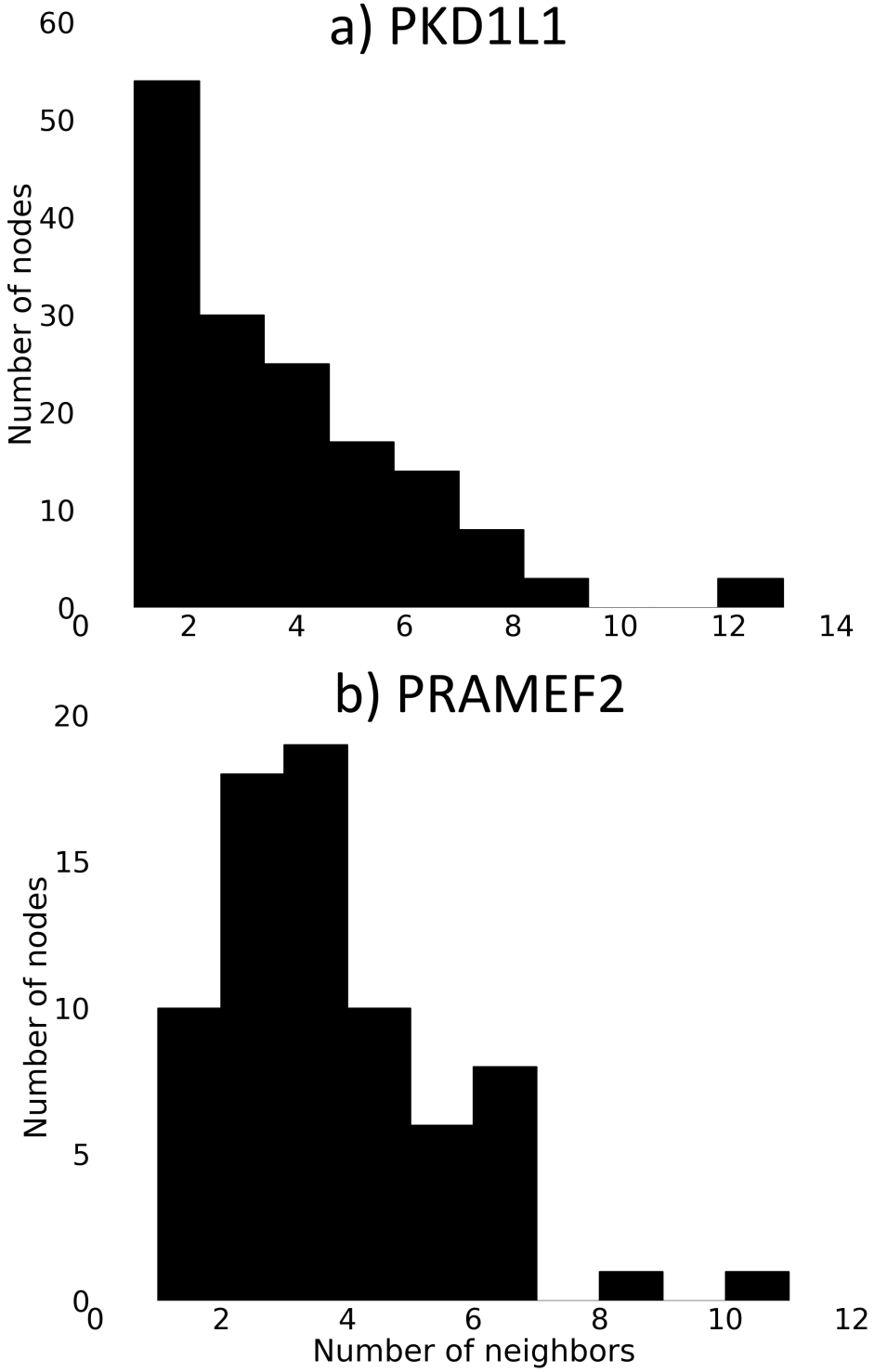


FIGURE S2.6: **Two examples for the distribution of the number of neighbors in the dominant component of networks with an excess of squares.** Most haplotype networks have a skewed distribution of the number of neighbors, of which the distribution in **a)** for *PKD1L1* (polycystic kidney disease 1 like 1) is representative. A minority of haplotype networks have a more symmetric distribution of this number of neighbors, as exemplified by **b)** for the network of **PRAMEF2** (PRAME family member 2).

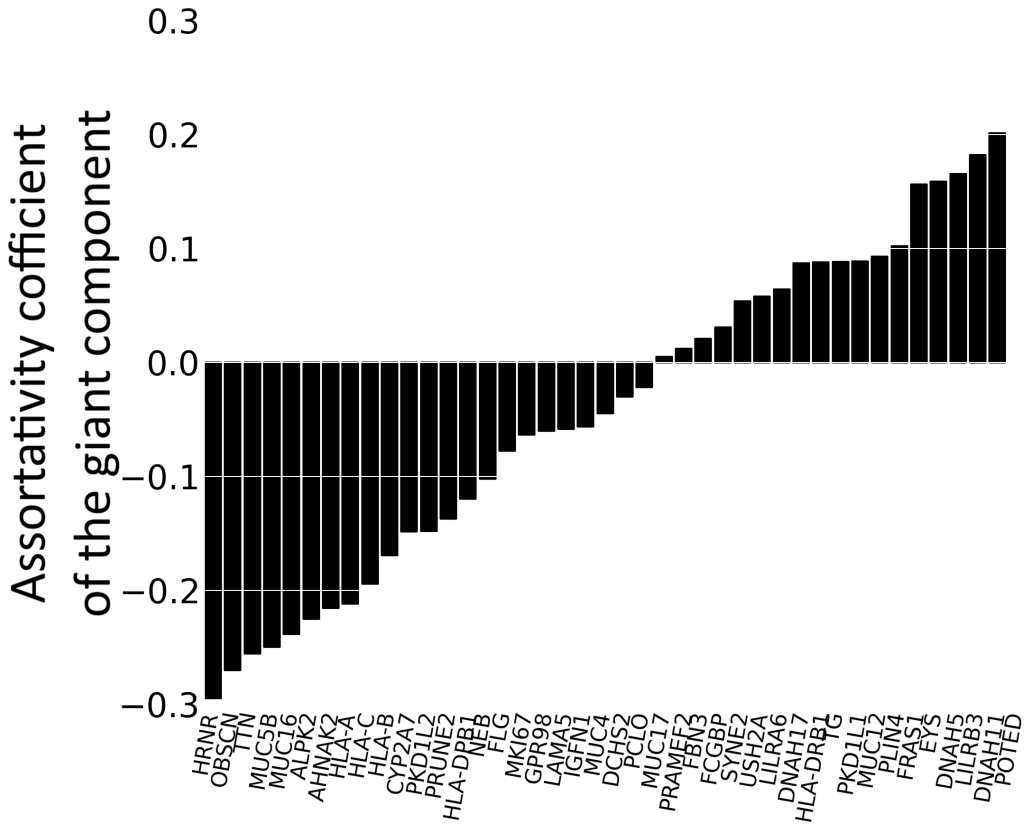


FIGURE S2.7: **Assortativity coefficient of haplotype networks of genes with an excess of squares.** A graph is (dis)assortative if nodes with many neighbors tend to connect with other nodes that have many (few) neighbors. This property can be quantified through an assortativity coefficient, which is the Pearson correlation coefficient of degrees between every pair of neighboring nodes [182]. The higher this assortativity coefficient, the higher the tendency of a node to connect to other nodes with similar number of neighbors. The graph shows the assortativity coefficient (vertical axis) for the largest component of the haplotype network of each gene with a significant excess of squares (horizontal axis).

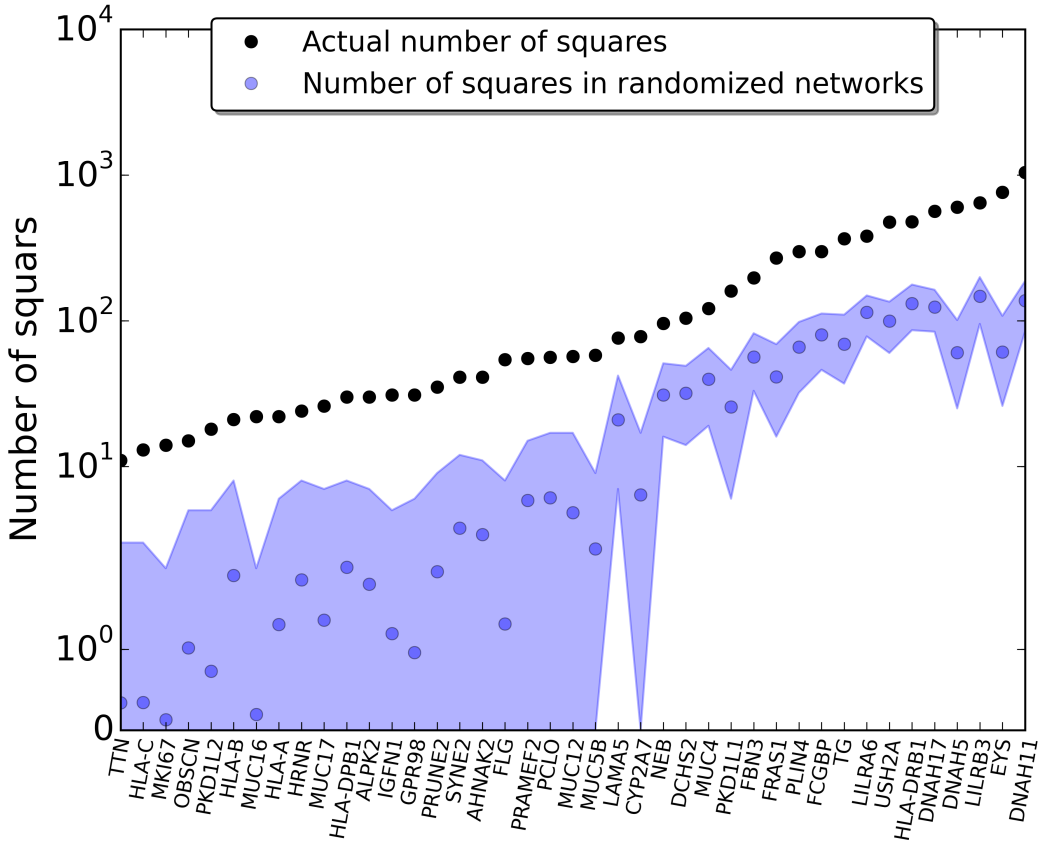


FIGURE S2.8: **Recombination cannot produce the observed number of squares.** For each of 41 genes with a significant excess of squares (horizontal axis), the vertical axis shows the number of squares in the largest components of the gene's haplotype network (black circles), and the mean number of squares for corresponding networks created through 1,000 population simulations with recombination (blue circles, see methods). The shaded area shows the minimum and maximum number of squares in 1,000 randomized networks for each gene. From the 42 genes with an excess of cycles, one gene (*POTED*, i.e., POTE ANKYRIN DOMAIN FAMILY, MEMBER D) was excluded from the analysis because it did not have any synonymous mutations, and so we could not estimate its recombination rate.

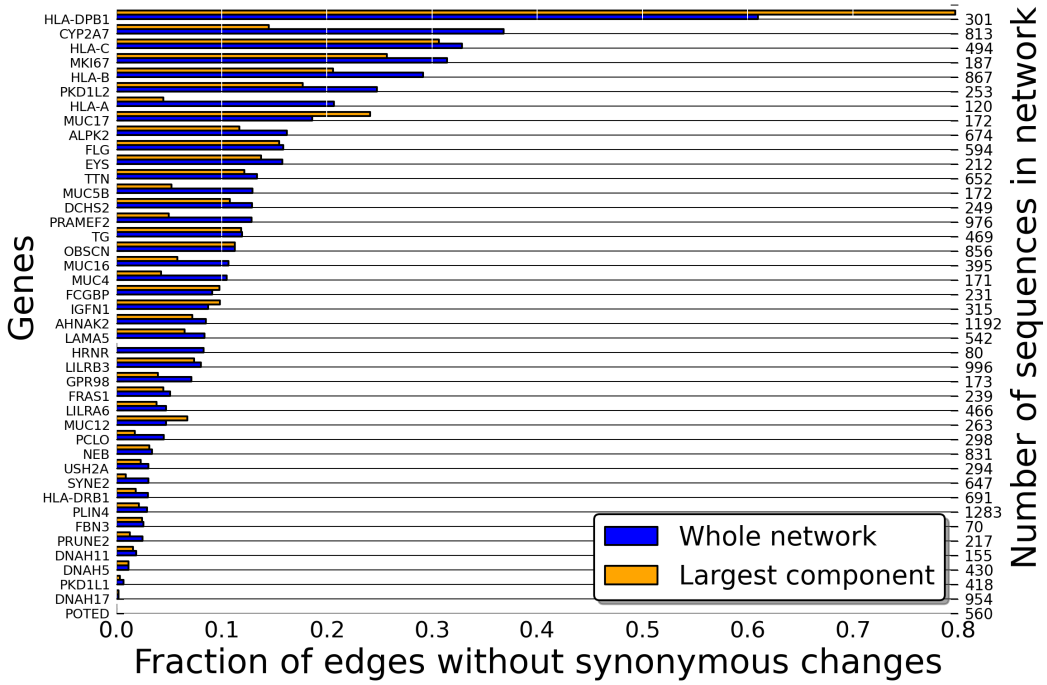


FIGURE S2.9: **Fraction of links without a single synonymous change (horizontal axis) in the dominant component and the whole haplotype network of those 42 genes (left vertical axis) with significantly more squares than expected by chance alone.** The numbers on the right vertical axis show the size of each haplotype network.

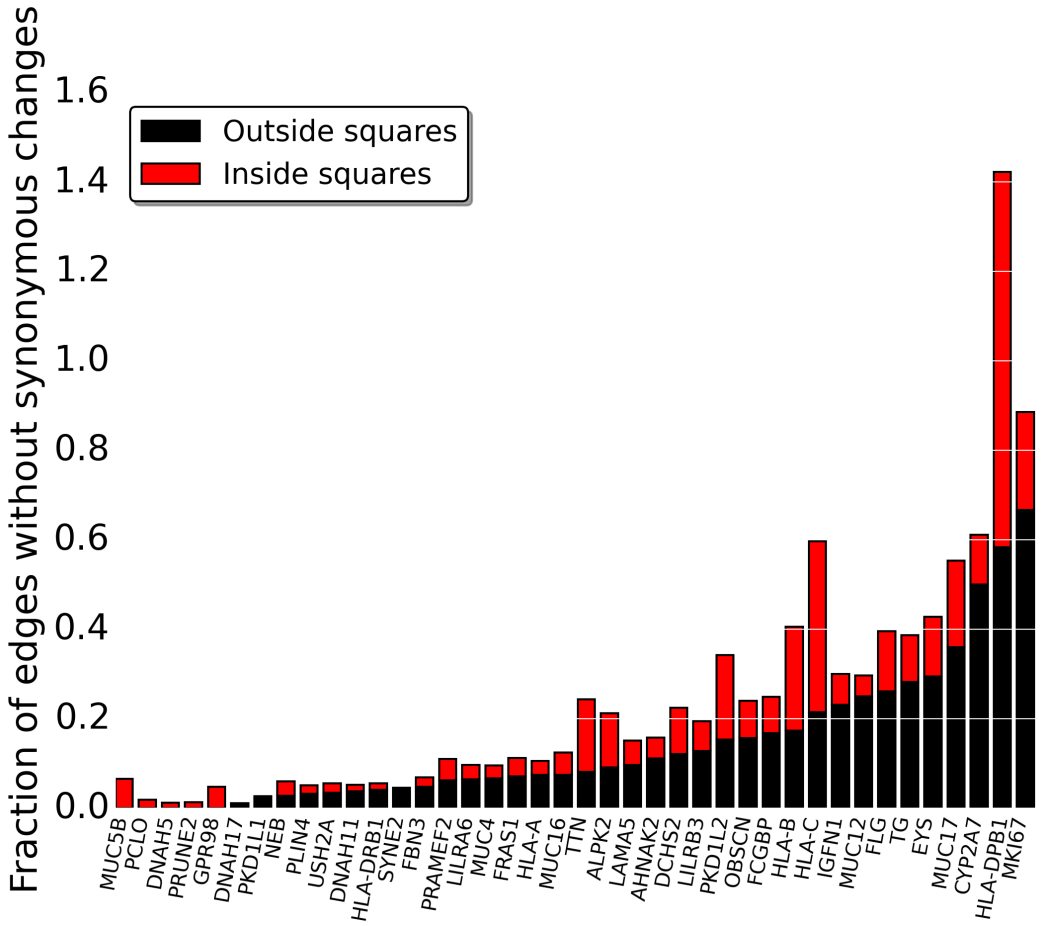


FIGURE S2.10: **The fraction of links without a single synonymous change inside and outside squares.** For each of 42 genes (horizontal axis) with significantly more squares than expected by chance alone, vertical bars show the fraction of links with no synonymous change for links that are part of a square (black bars) and that are not part of a square (red bars). The fraction of links without synonymous mutations is not significantly different for links inside squares compared to links outside squares for any gene (Mann-Whitney U test at $p = 0.05$ – corrected for multiple testing using [17]).

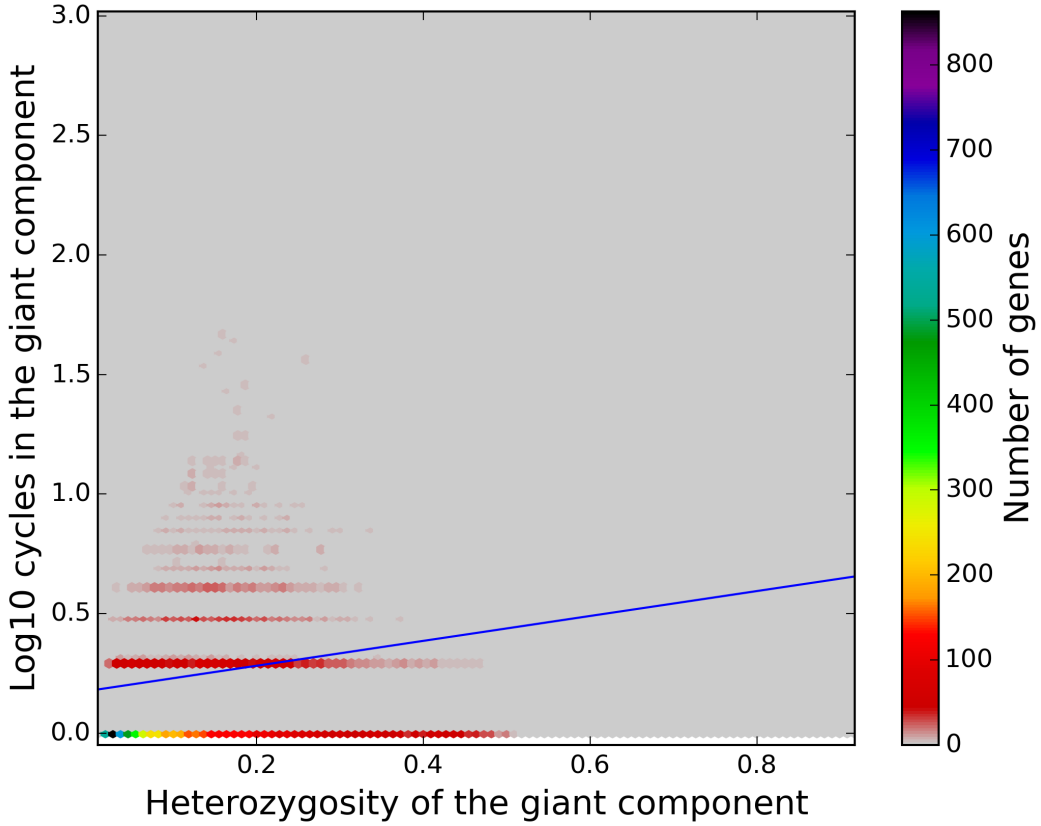


FIGURE S2.11: **Association between gene heterozygosity and number of squares in the dominant component of a gene's haplotype network.** We calculated the heterozygosity of each gene (n=12,235) as the average fraction of individuals heterozygous in that gene, where we took the average across all polymorphic sites in the population. The correlation is very weak but significant (Pearson's r= 0.066; $p = 3.42 \times 10^{-13}$; n=12,235). The blue line is based on linear regression.

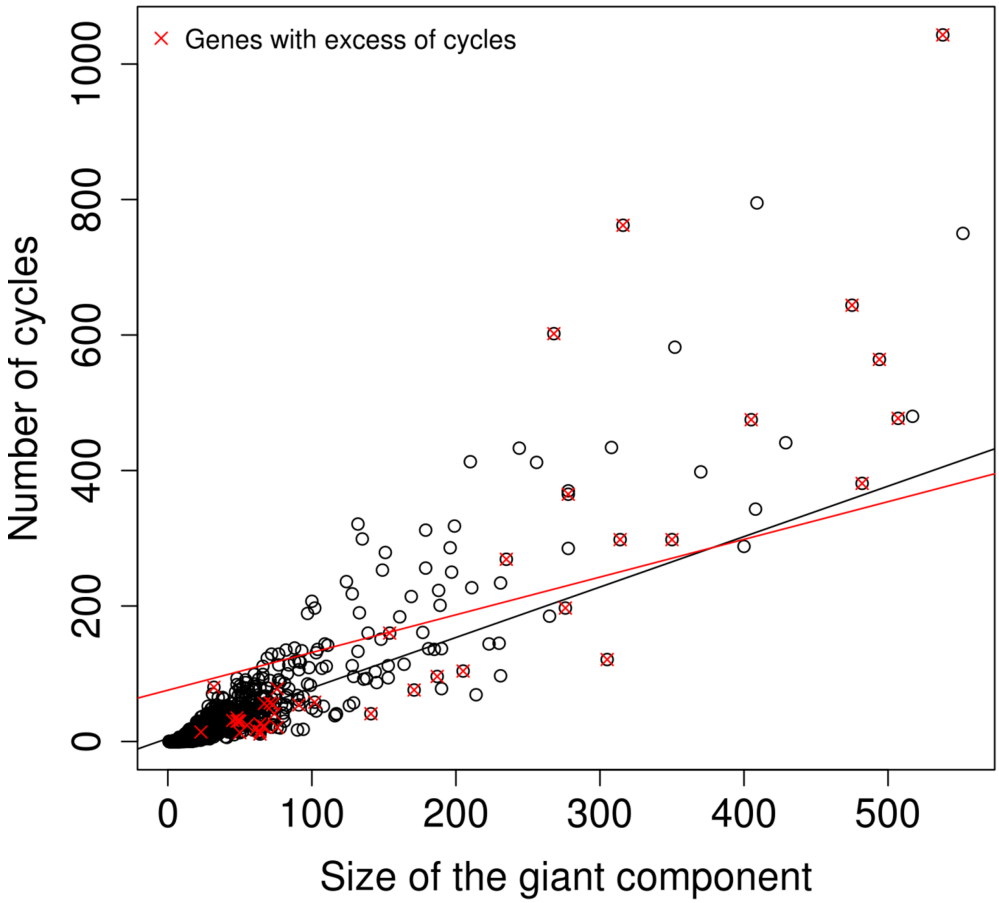


FIGURE S2.12: **Correlation between the size of the dominant component and the number of cycles in the dominant component of haplotype networks (based on 12235 genes).** Lines show results of linear regression analysis. Red specifies genes with a significant excess of cycles in their dominant component (42 genes). The size of the dominant component and the number of cycles are significantly correlated both across all genes and across genes with an excess of cycles (Pearson's product-moment correlation, p-value= $2.2 \times 10^{-16}$ and p-value= $1.04 \times 10^{-13}$ for all genes and for genes with an excess of cycles).

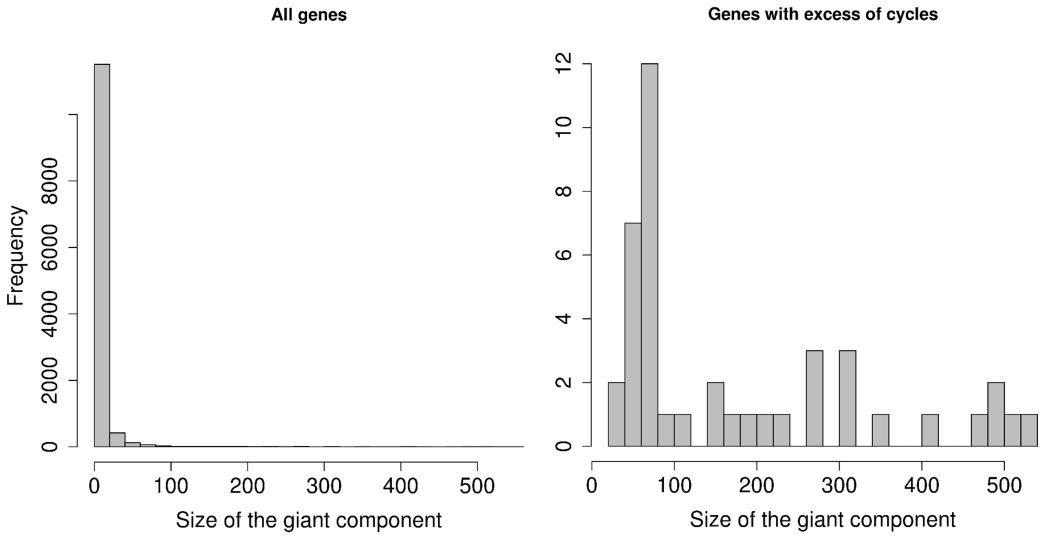


FIGURE S2.13: **Distribution of the size of the dominant component in gene haplotype networks.** The left panel shows this distribution for all 12,235 genes with at least one amino acid changing mutation (mean size of 35.7 haplotypes), and the right panel shows the distribution for those 42 genes with excess of cycles (mean size of 179.0 haplotypes). The two distributions are significantly different from each other (independent 2-group Mann- Whitney U Test, p-value= $2.2 \times 10^{-16}$).

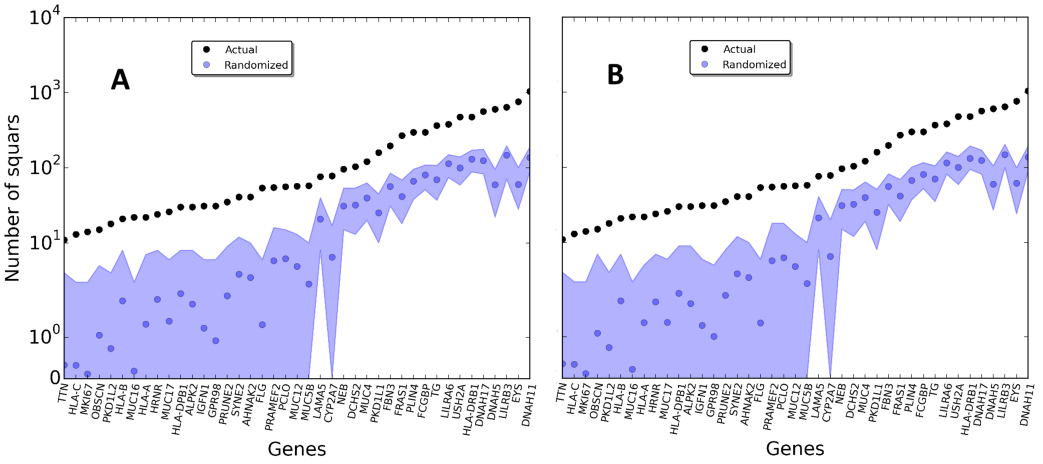


FIGURE S2.14: **Elevated recombination rates or increased effective population size cannot explain the observed number of cycles.** The vertical axes show the number of squares in the largest components of a gene's haplotype network (black circles), and the mean number of squares for corresponding networks created through 1,000 population simulations with recombination (blue circles, see methods). The shaded areas show the minimum and maximum number of squares in 1,000 randomized networks for each gene. **a)** Randomized networks were constructed with twice the average recombination rate than in the human genes, i.e. 1.90 $cM/Mb$. **b)** Randomized networks were constructed based on ten times the estimated effective population size of humans, i.e. 100,000 individuals. All other calculations and procedures are the same as described in the methods section describing how randomized networks with recombination were generated. From the 42 genes with an excess of cycles, one gene (*POTED*) was excluded from the analysis because it did not have any synonymous mutations, and so we could not estimate its recombination rate.

TABLE S2.1: **Genes that showed a signal of positive selection in the XP-CLR (cross-population composite likelihood ratio) test [37].** Column one shows gene names and column two show the population pairs in which the gene was identified as significant. Numbers in front of population pairs show the p-value of the most significant test statistic window overlapping the gene. CEU: Utah Residents with Northern and Western European ancestry; CHB: Han Chinese in Beijing, China;YRI: Yoruba in Ibadan, Nigeria.

| **Gene name** | **Population pair** |
|---|---|
| NEB | CEU compared against YRI |
| IGFN1 | YRI compared against CEU |
| FLG | YRI compared against CEU |
| PKD1L1 | YRI compared against CEU |
| GPR98 | CEU compared against CHB |
| FRAS1 | YRI compared against CHB |

## 2.8 Supplementary tables

TABLE S2.2: **Genes under positive selection as detected from Selectome database [209].** The Selectome database computes $d_N/d_S$ ratio on branches of the phylogenetic tree of vertebrates and, after correcting for multiple testing, identifies genes that have a $d_N/d_S$ ratio exceeding one on any specific tree branch. The table shows genes among the 42 genes with excess of squares in their network's dominant component that were detected by Selectome to be under positive selection. The second column shows that branch on which the gene was detected to be under positive selection.

| **Gene** | **Taxon** |
|---|---|
| DNAH5 | Euteleostomi |
| USH2A | Euteleostomi |
| PRAMEF2 | Primates |
| LAMA5 | Euteleostomi |
| FRAS1 | Euteleostomi |
| FBN3 | Euteleostomi |
| IGFN1 | Euteleostomi |
| GPR98 | Euteleostomi |
| PCLO | Euteleostomi |
| DNAH17 | Euteleostomi |
| HLA-DRB1 | Primates |
| FCGBP | Euteleostomi |

TABLE S2.3: **Hypergeometric test on genes under positive selection according to the XP-CLR test.** From a total of six possible tests between pairs of genes in the three populations (YRI, CEU and CHB), only four tests showed evidence that any of the 42 genes with an excess of cycles were under positive selection. Table columns, from left to right, show the corresponding population pairs, the total number of genes in the analysis, the number of genes under positive selection according to the test, the number of genes under positive selection among the 42 genes with an excess of cycles, and the p-value of the hypergeometric test. A p-value lower than 0.01 indicates that it is unlikely to find as many genes in our dataset to be under positive selection by chance alone.

| **Population** | **Genes in the population** | **Genes under positive selection in the population** | **Genes under positive selection that are part of our dataset** | **Hypergeometric test p-value** |
|---|---|---|---|---|
| CEU – YRI | 19221 | 371 | 1 | 0.19 |
| YRI – CEU | 19221 | 408 | 3 | 0.01 |
| CEU – CHB | 19221 | 342 | 1 | 0.17 |
| YRI – CHB | 19221 | 375 | 1 | 0.20 |

# Chapter 3

# Effect of population size and mutation rate on the evolution of RNA sequences on an adaptive landscape determined by RNA folding

Ali R. Vahdati, Kathleen Sprouffske, Andreas Wagner

# *Abstract*

The dynamics of populations evolving on an adaptive landscape depends on multiple factors, including the structure of the landscape, the rate of mutations, and effective population size. Existing theoretical work often makes ad hoc and simplifying assumptions about landscape structure, whereas experimental work can vary important parameters only to a limited extent. We here overcome some of these limitations by simulating the adaptive evolution of RNA molecules, whose fitness is determined by the thermodynamics of RNA secondary structure folding. We study the influence of mutation rates and populations sizes on final mean population fitness, on the substitution rates of mutations, and on population diversity. We show that evolutionary dynamics cannot be understood as a function of mutation rate $\mu$, population size $N$, or population mutation rate $N\mu$ alone. For example, at a given mutation rate, clonal interference prevents the fixation of beneficial mutations as population size increases, but larger populations still arrive at a higher mean fitness. In addition, at the highest population mutation rates we study, mean final fitness increases with population size, because small populations are driven to low fitness by the relatively higher incidence of mutations they experience. Our observations show that mutation rate and population size can interact in complex ways to influence the adaptive dynamics of a population on a biophysically motivated fitness landscape.

## 3.1 Background

Perhaps the most fundamental process in Darwinian evolution is a population's exploration of an adaptive landscape [271] by mutation and selection. As a population scales ever higher peaks in such a landscape, its mean fitness increases. (A fitness peak refers to one or more sequences with higher fitness than all their neighbors.) Many factors influence this process. Among them is the structure of the landscape itself, including its number of peaks, environmental changes that might influence this structure, the presence and incidence of recombination, the rate of DNA mutations, the kinds of genetic changes that such mutations cause, and population size [53, 54, 93, 110, 153, 166, 194, 263]. To understand these factors and how they interact to affect adaptive evolution is not just of academic interest. It may also help predict the outcome of adaptive evolution, for example in pathogens and their arms races with human and non-human hosts [79, 137, 148, 235, 249].

Unfortunately, the factors influencing adaptive evolution interact in complex ways. Here we focus on two such factors, mutations and their rate, as well as the effective size of a population $N_e$ [34, 154]. We study how these factors interact in the adaptive evolution of RNA molecules subject to mutation and selection on an unchanging fitness landscape.

Both separately and jointly, the two factors influence adaptive evolution in complex ways. Consider population size. On the one hand, adaptive evolution may be more rapid in large populations. First, larger populations produce more mutant individuals per generation, which helps explore more genotypes and find optimal genotypes faster than smaller populations. Second, natural selection is more effective in larger populations [190]. Specifically, as effective population size $N_e$ increases, natural selection becomes more effective in fixing beneficial mutations and removing deleterious mutations. In other words, the substitution rate of beneficial mutations is an increasing function of $N_e$, and the substitution rate of deleterious mutations a decreasing function of $N_e$ [3, 139]. Third, if mutation rates and population sizes are large enough, then some individuals in large populations will experience double mutations that can help them cross fitness valleys and explore genotypes that would otherwise be inaccessible [235], a phenomenon also known as stochastic tunneling [4, 108, 128, 259, 262].

On the other hand, there are also reasons why adaptive evolution may be

more rapid in smaller populations. First, such populations experience little or no clonal interference, a phenomenon that can slow down the adaptation rate in large and polymorphic populations [81, 235]. In clonal interference, multiple beneficial mutations coexist in a population at the same time. In the absence of recombination, individuals harboring different beneficial mutations compete with each other, which can slow down the fixation of beneficial mutations and thus adaptive evolution. Second, small populations experience stronger genetic drift and the stochastic changes in allele frequencies that can help a population cross a fitness valley [93, 110]. A different perspective on the same phenomenon is provided by considering the adaptive peaks in a multi-peaked adaptive landscape. Because only differences in fitness effects that are greater than the reciprocal of the population size ($1/N_e$) are visible to selection [190], some fitness peaks separated by a valley will merge as population size decreases, thus reducing the number of peaks in the landscape [110, 137, 235]. This will decrease the likelihood that a population becomes trapped on a local peak, and increase its chances to find the landscape's global fitness peak.

Further complications ensue if one considers the influence of mutations and the distribution of their fitness effects [46, 68]. These effects fall into three broad categories, deleterious, neutral, and beneficial. While the fate of neutral mutations is independent of population size [3, 190], this does no longer hold for beneficial or deleterious mutations. To be sure, strongly deleterious (lethal) mutations get eliminated rapidly, and strongly beneficial mutations sweep to fixation rapidly, but the fate of weakly deleterious and weakly beneficial mutations can depend on stochastic events caused by genetic drift and thus on population size. For example, weakly deleterious mutations can persist for substantial amounts of time, or even become fixed in small populations.

As a result of these interactions between mutation rate and population size, the substitution rate of mutations is expected to show a U-shaped relationship with $N_e$ [139]. That is, at small $N_e$, many slightly deleterious mutations become fixed. At large $N_e$, many slightly beneficial mutations become fixed, because positive selection is strong. At intermediate $N_e$, fewer mutations become fixed. The exact form of this relationship, however, depends strongly on the distribution of mutational fitness effects [46, 68, 239].

Existing work to elucidate the role of population size and mutation rate on

adaptive dynamics falls into two categories. The first comprises computational and theoretical studies to understand these dynamics [29, 53, 54, 123, 137, 153]. Because they do not use data from empirical adaptive landscapes, such studies usually make ad hoc assumptions about the structure of a fitness landscapes, the fitness effects of individual mutations, non-additive (epistatic) interactions of mutations [45, 248], and so on. Violations of these assumptions may affect the evolutionary dynamics [139]. For example, the effective population size $N_e$ and the substitution rate of beneficial mutations are expected to show a positive association if beneficial mutations are rare [139]. However, the incidence of beneficial mutations may change when the environment changes, or while a population explores a fitness landscape. Such change can affect the substitution rate of beneficial mutations, and thus also the rate of adaptive evolution.

Other studies use experimental approaches. Unlike theoretical studies, they examine fitness landscapes of realistic complexity. However, because such landscapes are very large and may involve astronomically many genotypes, we usually have very limited knowledge about the structure of these landscapes and about a population's evolutionary trajectories on them [135, 216]. Moreover, experimental studies are subject to limited replication, and can thus vary mutation rates, population sizes, and other relevant parameters only to a limited extent.

Here we overcome some of these limitations by simulating adaptive evolution on a biophysically motivated adaptive landscape that does not require ad hoc assumptions about landscape structure. It is a landscape whose structure is determined by the thermodynamics of RNA folding [223–225]. RNA molecules fold into secondary structures by internal pairing of complementary base pairs (G-C, A-U). Driven by thermal motions, an RNA molecule can fold and re-fold incessantly and thus adopt a spectrum of different secondary structures that differ in their free energy. The structure in which a molecule spends most of its time is the minimum free energy (MFE) structure [223, 273]. In our simulations, we use the fraction of time a molecule spends in a given fold –– the stability of this fold –– as a measure of fitness. This stability may itself be subject to selection [175]. A potential example is the stability of yeast mRNA secondary structures, which increases with gene expression levels [284]. For reasons of tractability, and considering existing precedents in modeling RNA evolution [5, 74, 224, 225], we assume that selection acts only on the stability of a single structure, but note that in nature a balance

between multiple secondary structures may be important [230, 240, 282].

Aside from using a biophysically motivated adaptive landscape, our simulation model also has the advantage that it does not require us to make ad hoc assumptions about fitness effects of mutations or about epistatic interactions of mutations, because these quantities are determined by the thermodynamics of folding. And with a simulation model, we can explore a wider range of mutation rates and population sizes than in experimental work. Although one might naively assume that evolutionary dynamics can be understood as a function of mutation rate $\mu$ or population mutation rate ($N\mu$) alone, our observations show otherwise.

## 3.2 Results

### 3.2.1 Short RNA sequences folding into any secondary structure are highly connected

Our evolution simulations build on two different kinds of RNA sequences. The first comprise all of those $4^{10} = (1,048,576)$ ten-nucleotide-long sequences that fold into some secondary structure in their minimum free energy (MFE) state. Before studying the evolutionary dynamics of these molecules, we first characterized how they are organized in RNA genotype space. To this end, we first determined by exhaustive enumeration that there are 39,410 sequences (3.76% of sequence space) with some MFE secondary structure, and that they form nine distinct secondary structures. Each of these structures has a single stem-loop but with different nucleotides involved in the stem (Table 3.1). Although these sequences comprise a small fraction of the whole genotype space, they are highly accessible from one another through single mutations. This can be shown by constructing a genotype network, i.e., a graph whose nodes are sequences that form some secondary structure (regardless of the identity of that structure), and whose edges connect two sequences that differ by a single point mutation. This graph has five connected components. (A component is a set of nodes that are accessible from each other through a path of one or more edges.) However, one of these components contains the vast majority (99.24%, 39,109) of sequences (Figure 3.1).

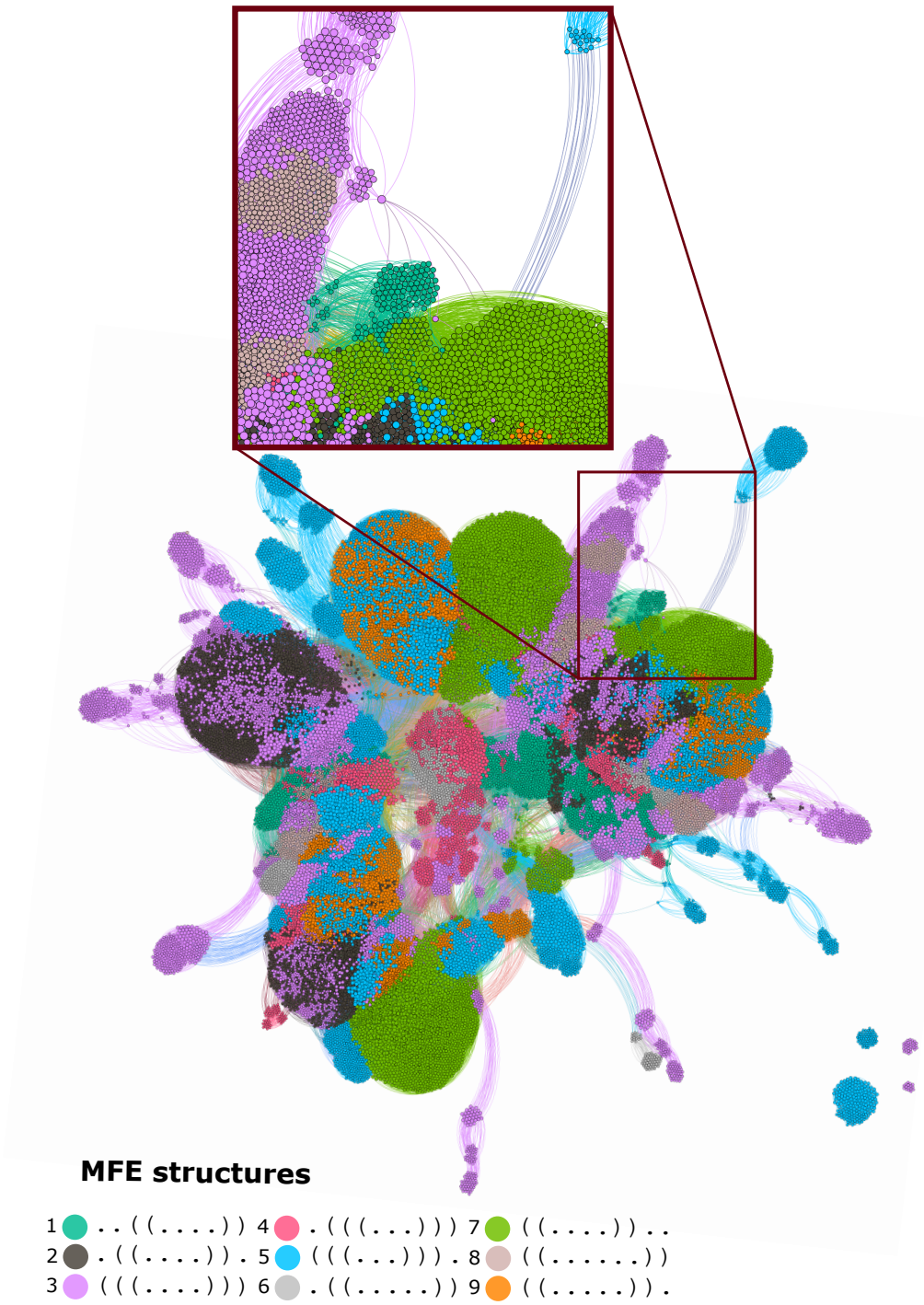


FIGURE 3.1: **The genotype network of RNA sequences of length 10.** Each circle (node) corresponds to a sequence. Two nodes are connected if they differ by a single point mutation. Nodes with the same color have the same minimum free energy secondary structure (Legend).The inset enlarges a part of the largest component. Nodes are clustered based on their number of shared connections (based on ForceAtlas2 embedding in Gephi [16]). For clarity of representation, our display allows for overlapping nodes, such that the actual number of nodes may be more than the number of nodes that are visible. The graph in the figure illustrates the intertwined organization of different genotype networks and genotype sets. Because of its large number of nodes (39401) and edges (311000), not all nodes and edges are visible, and accurate accounting of component numbers is thus not possible.

One can subdivide the nodes (sequences) in this graph into subsets of sequences associated with each one of the nine MFE secondary structures. Each such subset itself forms a genotype network with multiple connected components. Specifically, depending on the structure, these networks comprise between 943 to 8,513 nodes, and have between 3 to 21 connected components each. All of them are positively assortative, with assortativity values between 0.13 and 0.82 (see Methods), meaning that highly connected sequences tend to be connected to other highly connected sequences. It takes 5 to 10 mutations to travel between the most distant two nodes while staying within the largest component of each network (see column "Diameter" in Table 3.1).

TABLE 3.1: **Properties of genotype networks of RNA molecules of length 10 that fold into the nine possible secondary structures.** Columns from left to right: 'ID' : an identifier for the secondary structure; 'Vertices' : number of sequences folding into the structure; 'GC vertices' : number of edges in the dominant component of the genotype network formed by the sequences; 'Components' : number of connected components within each network (a connected component is a set of sequences which are all accessible from each other through a series of single point mutations that preserve the structure); 'Assortativity' : assortativity coefficient of the largest connected component. The assortativity coefficient indicates to what extent sequences have neighbors with degrees (numbers of neighbors) similar to themselves [13]; 'Diameter' : the diameter of the largest connected component. The diameter of a network is the largest minimal distance between any pair of nodes in a connected component; 'Structure' : MFE structure of the sequences in the network; 'Min-Max time in MFE structure' : range of the fraction of times that sequences folding into the MFE structure spend in this structure. More time spent in a structure corresponds to higher fitness in our model.

| ID | Vertices | GC vertices | Components | Assortativity | Diameter | Structure | Min-Max time in MFE structure |
|---|---|---|---|---|---|---|---|
| Str1 | 1,728 | 731 | 3 | 0.75 | 9 | ..((....)) | 0.36-0.85 |
| Str2 | 5,717 | 2,445 | 6 | 0.70 | 11 | .((....)). | 0.38-0.98 |
| Str3 | 7,790 | 1,487 | 13 | 0.79 | 11 | (((....))) | 0.42-0.97 |
| Str4 | 2,286 | 506 | 10 | 0.67 | 6 | .(((...))) | 0.46-0.90 |
| Str5 | 6,934 | 1,335 | 21 | 0.82 | 10 | (((...))). | 0.52-0.98 |
| Str6 | 943 | 384 | 5 | 0.55 | 6 | .((.....)) | 0.40-0.64 |
| Str7 | 7,765 | 3,328 | 9 | 0.65 | 8 | ((....)).. | 0.38-0.98 |
| Str8 | 1,437 | 475 | 5 | 0.13 | 6 | ((.....)) | 0.39-0.64 |
| Str9 | 4,801 | 2,115 | 4 | 0.58 | 8 | ((....)). | 0.39-0.95 |

Our simulations of evolving populations use the fraction of time that sequences spend in their MFE structure as a measure of fitness. This fraction varies, depending on structure, between 0.27 and 0.97 among the nine structures. Here, a value of 0.27 (0.97) means that a sequence spends 27 percent of the time in its MFE structure, and the remaining 73 (3) percent in some other structures with higher free energy. (The MFE structure can be viewed as the structure in which a sequence spends more time than in any other structure, even though it may not spend the majority of its time in this structure.) Within the genotype network of each structure, it varies between values ranging from 0.27 to 0.96 for structure "`.((....))`." to values ranging from 0.51 to 0.71 for structure "`.((.....))`".

How an evolving population explores a fitness landscape depends in part on the fraction of its sequences' neighbors that are neutral. If a population has a larger neutral neighborhood, it may be able to access larger regions of the landscape through non-deleterious mutations, and may have a higher chance of finding beneficial mutations and new phenotypes. We computed the size of neutral neighborhoods, because it may be important for our evolutionary analysis. This size is a function of effective population size $N_e$ [96], which in our case is identical to the census population size $N$, because the populations we simulate are unstructured, do not experience migration, and do not fluctuate in size. Following standard population genetic theory [124, 191], we consider two neighboring sequences neutral if their fitness differs by less than $1/N$. Figure S3.1a shows neutral neighborhood size as an average over 1,000 randomly sampled RNA molecules of length 10 that fold into one of the nine structures we consider (Table 3.1). Unsurprisingly, neutral neighborhood size decreases with increasing population size, where neutral evolution and crossing of fitness valleys becomes more difficult.

To ensure that any observations we obtain from our simulations are not artefacts of using very short and non-biological sequences, we also simulated the evolution of four longer biological RNA molecules (30-43nts) that originate from different organisms, have different functions, and fold into different predicted secondary structures (Table 3.2). Specifically, these sequences include a ribozyme, a noncoding transcript, a small non-messenger RNA (snmRNA), and a small nuclear RNA (snoRNA). While the large number of sequences folding into such longer structures [225] precludes an exhaustive analysis of their genotype networks, we find that the neutral neighborhoods of these genotype networks also decrease in size with increasing population

size (Figure S3.1b).

We quantified the ruggedness of the fitness landscapes of our RNA molecules in two ways. First, we counted the number of fitness peaks in each landscape of sequences of length 10, where we define a fitness peak as one or more sequences whose neighbors all have lower fitness. With the exception of structure 2 (Str2) and structure 3 (Str3), which have 10 and 23 peaks, respectively, all structures have fewer than 10 peaks (Figure S3.2). This analysis was not possible for the biological sequences, where too many sequences fold into any one structure. Second, we estimated the incidence of reciprocal sign epistasis, which causes fitness valleys to exist between a sequence and its two-mutant neighbor. In epistasis, the fitness effect of an allele depends on other alleles. Sign epistasis occurs when the sign of the fitness effect of an allele changes (e.g. from beneficial to deleterious) due to epistatic interactions. When a sequence and its two-mutant neighbor both show higher fitness than the two single-mutants connecting them in sequence space, one speaks of reciprocal sign epistasis [205]. We find that fewer than 10 percent of such sequence quadruplets show reciprocal sign epistasis. This holds regardless of whether we consider sequences of length 10 or biological sequences (Figure S3.3). Overall, these analysis show that the landscapes we examine are not highly rugged.

We simulated the adaptive evolution of sequences forming each one of the nine secondary structures of length 10, as well as each one of the four biological sequences. That is, we evolved populations of such sequences through 800 cycles (generations) of mutation and selection favoring an increase in the time that a sequence spends in the focal secondary structure (see Methods). We performed 50 replicates for each population simulation. Because we were interested in the influence of population size $N$ and mutation rate $\mu$ on the speed of adaptive evolution, we varied both parameters systematically ($0.0001 < \mu < 1, 0.01 < N\mu < 10$). In the following, we find it most useful to analyze our observations separately for varying $\mu$ and varying population mutation rates $N\mu$.

TABLE 3.2: **Biological RNA sequences used in this study.** Columns from left to right: 'Identifier' : the fRNAdb database identifiers [127] for the four sequences considered here; 'Organism' : the organism in which the RNA sequence was identified; 'RNA type' : functional classification of the RNA sequence; 'Sequence' : the sequence of the RNA; 'Secondary structure': the secondary structure of the RNA sequence. We computed secondary structures using the `fold` function from the ViennaRNA package [150].

| Identifier | Organism | RNA type | Sequence | Secondary structure |
|---|---|---|---|---|
| AF357483 | Mus musculus | snmRNA | AAGCAAUUGUUUUACUUACAGUCUGGAGAA | ...(((((((.......))))))..))..... |
| Z71666 | Saccharomyces cerevisiae | snoRNA | AGGCGUGUAACAUUUAUUGGUUACAACAUG | .....((((((........))))))..... |
| AB055777 | Homo sapiens | noncoding transcript | CUCUUUUACCAAGGACCCGCCAACAUGGGC | .(((((....)))))((((......)))). |
| AF036740 | Schistosoma mansoni | ribozyme | AUCCAGCUCACGAGUCCCAAAUAGGACGAAACGCGUCCUCCAU | ......(((((((.....))))))).... |

### 3.2.2 Adaptive evolution under varying mutation rate $\mu$

$\mu = 0.0001$

At this low mutation rate $N\mu << 1$ for all population sizes we considered. All populations of sequences with length 10 reach similar mean fitness at the end of evolution (Figure 3.2a), except for a minority of structures where the largest populations reaches a significantly higher mean fitness (Str2, Str7 and Str9, Figure S3.4). In contrast, biological sequences show a consistent and significant increase in final mean fitness as population size increases (Figure 3.3a). The likely reason of this difference between sequences of length 10 and biological sequences is that the incidence of neutral, beneficial, and deleterious mutations differs between them. In sequences of length 10, beneficial mutations are less common than deleterious ones, whereas in biological sequences, they are more common (Figure S3.5). These differences may result from differences in landscape size. Our biological sequences have a vastly larger landscape ($4^{30}$-$4^{43}$ sequences) than sequences of length 10 ($4^{10}$ sequences), which may influence the distribution of fitness effects. An additional difference may come from how we implemented selection. In sequences of length 10, we allowed only sequences whose MFE secondary structure matches the target structure to survive, which permitted us to restrict the evolutionary dynamics to sequences with the same MFE structure. In contrast, for biological sequences, we allowed any sequence that folds into a given target structure to survive. Moreover, we initialized populations of biological sequences from random sequences whose fitness was less than 0.01, whereas populations of length 10 sequences started from sequences with a fitness in the bottom 5%. This is because biological sequence landscapes were too large to analyze exhaustively. These two is differences may also affect the distribution of fitness effects and consequently, the prevalence of beneficial mutations between the 10-nucleotide and biological sequences. As a result of the greater incidence of beneficial mutations, larger populations of biological sequences can increase their fitness more easily. It may seem surprising that population size makes a difference at mutation rates this small, but larger populations have an advantage at several levels. Firstly, in every generation, larger populations are slightly more diverse (Figures 3.2c and S3.6a), even though the difference between larger and smaller populations is minute. Second, across all 50 simulation replicates, larger populations visit more unique sequences than smaller populations (Figures 3.2b and S3.6b). In

other words, because larger populations produce more mutations per generation than smaller populations, they are better at exploring genotype space. Third, and consistent with this observation, larger populations also experience more nucleotide substitutions (Figure 3.2d), the majority of which are beneficial (e.g. Figure 3.2d). The reason is that selection is more efficient in larger populations [110, 137, 235]. The difference between sequences of length 10 and biological sequences highlights the importance of the distribution of mutational effects and of its interactions with population size for adaptation. When deleterious mutations are prevalent, larger populations may not adapt faster. However, when beneficial mutations are prevalent, larger populations may adapt significantly faster.

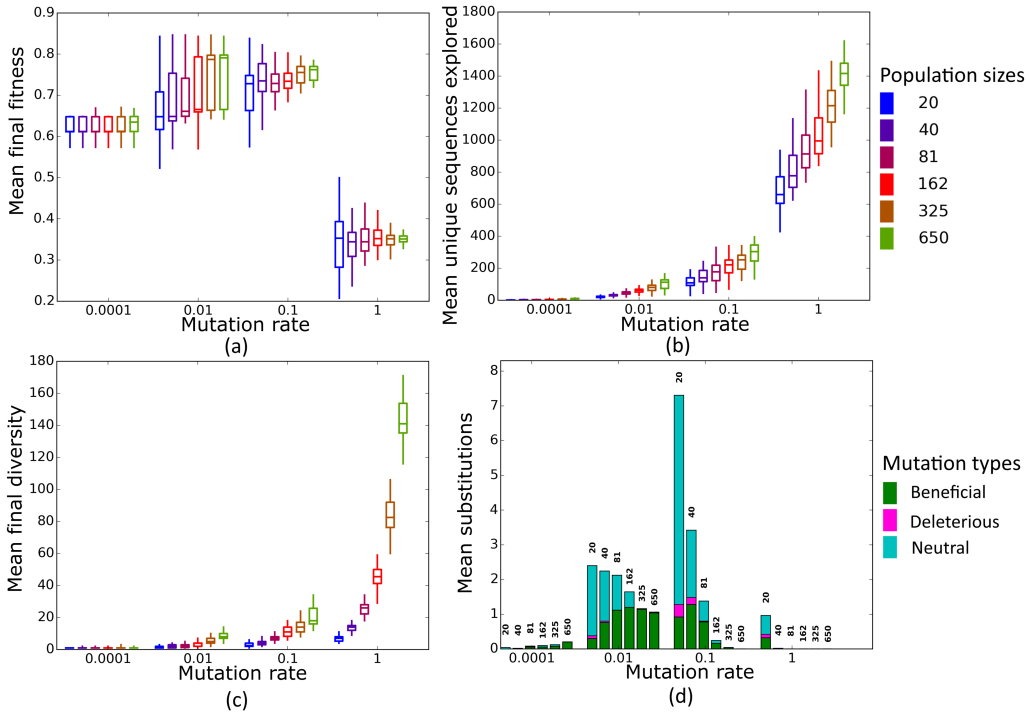


FIGURE 3.2: **Simulated evolution of sequences with secondary structure 1 (Str1, Table 3.1) at varying mutation rates and population sizes.** We randomly-selected a low-fitness sequence to initialize each simulation, and then simulated 800 generations of mutation and selection. We performed 50 replicate simulations for each mutation rate (horizontal axes) and population size (see Methods). Boxplots show **(a)** final mean population fitness, **(b)** total unique sequences explored, and **(c)** final population diversity (number of unique sequences at generation 800). Each box encloses the second and third quartiles of the 50 replicates, the center line corresponds to the median, and whiskers depict the minimum and maximum values obtained from any replicate, excluding the outliers. **(d)** Mean numbers of unique beneficial, deleterious, and neutral substitutions (green, pink, and cyan) are summarized as bars for the 50 replicates at each mutation rate (horizontal axis) and population size (labels above bars).

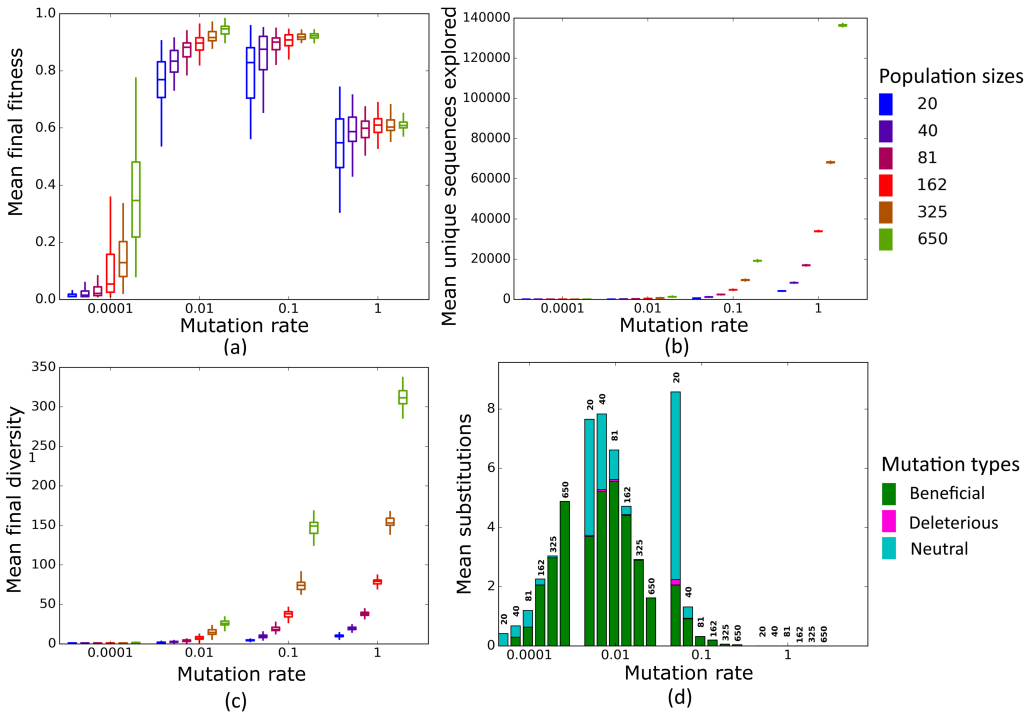


FIGURE 3.3: **Simulated evolution of sequences with secondary structure of AF036740 RNA sequence (Table 3.2) at varying mutation rates and population sizes.** We randomly-selected a low-fitness sequence to initialize each simulation, and then simulated 800 generations of mutation and selection. We performed 50 replicate simulations for each mutation rate (horizontal axes) and population size (see Methods). Boxplots show **(a)** final mean population fitness, **(b)** total unique sequences explored, and **(c)** final population diversity (number of unique sequences at generation 800). Each box encloses the second and third quartiles of the 50 replicates, the center line corresponds to the median, and the whiskers depict the minimum and maximum values obtained from any replicate, excluding the outliers. **(d)** Mean numbers of unique beneficial, deleterious, and neutral substitutions (green, pink, and cyan) are summarized as bars for the 50 replicates at each mutation rate (horizontal axis) and population size (labels above bars).

$\mu = 0.01$

At this mutation rate, all populations reach a higher final mean fitness than at $\mu = 0.0001$ (Figures 3.2a and 3.3a). Two different regimes are relevant to understand the evolutionary dynamics of populations at different sizes $N$. At smaller population sizes ($N = 20$, $N = 40$, and $N = 81$), $N\mu < 1$, whereas at larger sizes ($N = 162$, $N = 325$, and $N = 650$) $N\mu > 1$. In the latter case, populations are expected to be polymorphic most of the time ([198]), which raises the possibility of clonal interference. That is, a population may harbor more than one beneficial sequence variant, and the two sequences may compete for fixation, resulting in lower fixation rates for either variant. We first wished to find out whether clonal interference occurs in our populations. Figures S3.7a and S3.7b show the frequency of the average number of unique sequences per generation in each population, and classify these sequences according to their fitness effect -– beneficial, neutral, or deleterious -– relative to the ancestral sequence at the start of the simulation. Clearly, as $N$ increases, the number of unique beneficial alleles that are present at any one time in a population increases as well (Figures S3.7a and S3.7b). We also find that nucleotide substitution rates drop for populations with population mutation rates $N\mu > 1$ (i.e. $N = 162$, $N = 325$, and $N = 650$), both for sequences of length 10 (Figure 3.2d) and for biological sequences (Figure 3.3d). But despite increased clonal interference and decreased substitutions in large populations, we also find that larger populations generally have higher final mean fitness (Figure S3.8a). Specifically, final fitness is significantly higher for seven out of the nine structure of length 10 (all but Str4 and Str9), and for all biological sequences (Figure S3.8b). To find out what may be responsible for this increase, we pooled data from simulations at different population sizes, and asked whether final mean population fitness is correlated with two measures of population diversity, namely the total number of sequences explored by a population, and the total diversity of a population in the last generation (generation 800, see Methods). In populations of sequences of length 10, mean final population fitness showed a significant positive association with the total number of explored sequences (Table S3.2, Figure S3.9a), and a significantly positive association with population diversity for all structures except Str1 (Table S4.4, Figure S3.9b). Mean final fitness has a significant positive association with total number of explored sequences and population diversity for biological sequences (Figures

S3.10a and S3.10b). We note that larger populations explore more unique sequences during evolution (Figure 3.2b) and are on average more diverse in the last generation (Figure 3.2c). Taken together, these observations suggest an explanation for the consistently higher fitness in large populations: Such populations explore more sequences and thus have higher standing variation, which increases the prevalence of beneficial alleles (Figures S3.7c and S3.7d). A greater number of beneficial alleles, in turn, is associated with an increase in the average fitness of a population (Figures S3.11a and S3.11b), even when no mutations are fixed. In sum, the final mean fitness of a population is not completely determined by clonal interference, but also depends on a population's genetic diversity.

$\mu = 0.1$

At this mutation rate, populations arrive at a mean final fitness similar to that at $\mu = 0.01$ (Figures 3.2a and 3.3a). All population sizes are in the regime of $N\mu > 1$ where clonal interference occurs and becomes stronger in large populations. For all but four sequences of length 10 (Figure S3.12a), we no longer observe a significant increase in average population fitness as population size increases, but such an increase still exists for biological sequences (Figure S3.12b). To explain the observation that mean fitness does not decline in larger populations, even though clonal interference becomes stronger, it helps again to consider the incidence of nucleotide substitutions and population diversity. At $\mu = 0.1$, smaller populations fix more mutations than large populations, whereas large populations fix hardly any mutations (Figures 3.2d and 3.3d) due to clonal interference. However, not unexpectedly, larger populations again explore more unique sequences than smaller populations (Figure 3.2b). This reinforces the notion that increased sequence exploration can override the influence of clonal interference on final mean fitness. Populations with few substitutions but high diversity and more beneficial mutations (Figure S3.13) have a higher average fitness than sequences with lower diversity and exploration but more substitutions. The difference between sequences of length 10 (little increase in mean fitness at larger $N$) and biological sequences (larger increase in mean fitness) is consistent with this notion. For example, populations with size $N = 650$ and size $N = 20$ differ in mean fitness by approximately 10% for the biological structure AF036740, but only by about 5% for Str1 of length 10. The reason is that the total number of

explored sequences increases to a much greater extent between the smallest and largest population size in biological sequences (ca. 30-fold) than for sequences of length 10 (7-fold) (Figure S3.14, similar patterns exist between other structures (data not shown)).

$\mu = 1$

In this regime, all populations have $N\mu >> 1$. Just as for $\mu = 0.1$, we do not observe dramatic differences in final mean fitness as population sizes vary (Figures 3.2a and 3.3a). More strikingly, however, mean fitness at all population sizes is lower than at smaller mutation rates. The reason of this fitness decrease is the high fraction of mutant sequences per generation. Each individual sequence on average experiences one mutation per generation, which drives a population away from high-fitness sequences. Consequently, the mean fitness of the population fluctuates around a low value, and populations fix few mutations.

### 3.2.3 Adaptive evolution under varying population mutation rates $N\mu$

As the preceding observations showed, mutation rates interact with population sizes to influence adaptive evolution. We next wanted to find out whether the population mutation rate $N\mu$, a central quantity in population genetics, is sufficient to capture this interaction.

$N\mu = 0.01$ **to** $N\mu = 1$

At these low to moderate population mutation rates, mean population fitness does not depend on population size (Figures 3.4a and 3.5a), nor does the mean final diversity of populations (Figures 3.4c and 3.5c), which suggests that $N\mu$ may be sufficient to describe the evolutionary dynamics of populations. However, at least for $N\mu = 1$, the number of explored sequences decreases with population size $N$ (Figure 3.4b and 3.5b). The likely reason is that smaller populations have larger neutral neighborhoods (Figures S3.1a and S3.1b), which means that fewer mutations will be eliminated by natural selection, and more sequences can be explored through mutation. This is

also consistent with the observation that larger populations experience fewer nucleotide substitutions, especially of neutral mutations, at $N\mu = 1$ (Figures 3.4d and 3.5d). It can also be explained by the larger size of neutral neighborhoods at small $N$, which leads to more neutral mutations, and thus to more neutral substitution events. In sum, even though final mean fitness does not depend on $N$ for small to moderate $N\mu$, population diversity and substitution rates do depend on population size. $N\mu$ is thus not the only relevant parameter describing the evolutionary dynamics of our populations.

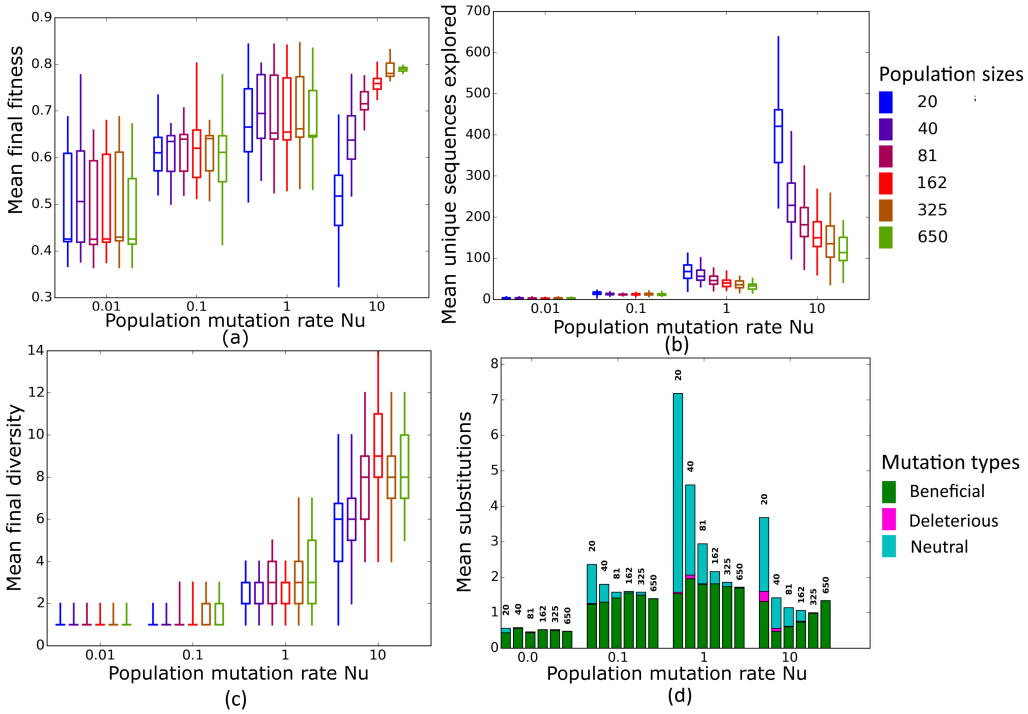


FIGURE 3.4: **Simulated evolution of sequences with secondary structure 1 (Str1, Table 3.1) at varying population mutation rates $N\mu$ and population sizes.** We randomly-selected a low-fitness sequence to initialize each simulation, and then simulated 800 generations of mutation and selection. We performed 50 replicate simulations for each mutation rate (horizontal axes) and population size (see Methods). Boxplots show **(a)** final mean population fitness, **(b)** total unique sequences explored, and **(c)** final population diversity (number of unique sequences at generation 800). Each box encloses the second and third quartiles of the 50 replicates, the center line corresponds to the median, and the whiskers depict the minimum and maximum values obtained from any replicate, excluding the outliers. **(d)** Mean numbers of unique beneficial, deleterious, and neutral substitutions (green, pink, and cyan) are summarized as bars for the 50 replicates at each mutation rate (horizontal axis) and population size (labels above bars).

### $N\mu = 10$

At the largest population mutation rates, $N$ affects not only the number of explored sequences (Figures 3.4c and 3.5c), the final population diversity (Figures 3.4b and 3.5b), and the number of substitution events (Figures 3.4d and

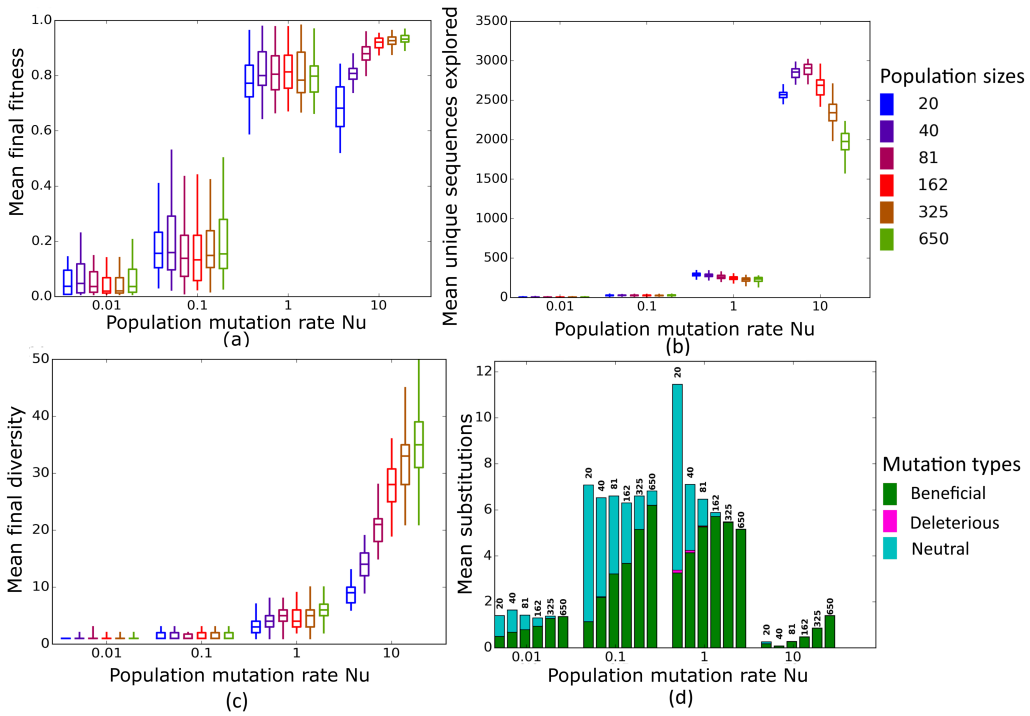


FIGURE 3.5: **Simulated evolution of sequences with secondary structure AF036740 RNA sequence (Table 3.2) at varying population mutation rates $N\mu$ and population sizes.** We randomly-selected a low-fitness sequence to initialize each simulation, and then simulated 800 generations of mutation and selection. We performed 50 replicate simulations for each mutation rate (horizontal axes) and population size (see Methods). Boxplots show **(a)** final mean population fitness, **(b)** total unique sequences explored, and **(c)** final population diversity (number of unique sequences at generation 800). Each box encloses the second and third quartiles of the 50 replicates, the center line corresponds to the median, and the whiskers depict the minimum and maximum values obtained from any replicate, excluding the outliers. **(d)** Mean numbers of unique beneficial, deleterious, and neutral substitutions (green, pink, and cyan) are summarized as bars for the 50 replicates at each mutation rate (horizontal axis) and population size (labels above bars).

3.5d), but also the final mean fitness (Figures 3.4a and 3.5a). This underscores that $N\mu$ cannot account for all aspects of the evolutionary dynamics. Specifically, at constant $N\mu = 10$, mean final fitness increases strongly with $N$ (Figures 3.4a and 3.5a). At least two causes can help explain this pattern. First, at constant $N\mu$, larger populations may fix more beneficial mutations, because selection is stronger in such populations. Second, and more importantly, a higher population mutation rate may be more destabilizing for smaller populations than for larger populations. For example, ten new mutations per population and generation means that half of all sequences in the smallest populations ($N = 20$) are mutated per generation, whereas only about 1.5 percent of sequences in the largest populations ($N = 650$) are mutated. Such a high incidence of mutation in the largest populations can drive a population away from a fitness peak, and overwhelm natural selection's power to increase mean fitness.

## 3.3 Discussion

Understanding the rate at which populations undergo evolutionary adaptation is central to research areas such as conservation biology ([76, 157, 234]), and microbial evolutionary biology ([14, 194, 247, 264]). Experimental approaches often have difficulties measuring quantities that are crucial to understand a population's evolutionary dynamics completely [64, 137, 141, 142], whereas theoretical approaches are often forced to make simplifying assumptions [29, 53, 54, 123, 153]). Here we tried to overcome some of these limitations by simulating the adaptive evolution of RNA molecules on a biophysically determined adaptive landscape. This helped us avoid making ad hoc assumptions about landscape structure, and allowed us to study adaptive dynamics in more detail than experimental approaches could. Our observations suggest an unexpectedly complex interaction between mutation rate and population size. First, at any one mutation rate, final population mean fitness tends to increase with population size, and especially for biological RNA sequences (Figure 3.3a). This holds even where $N\mu > 1$ and thus where clonal interference reduces the number of nucleotide substitutions. This observation is significant, because the substitution rate, especially that of beneficial mutations, is sometimes treated as being equivalent to the rate of adaptive evolution [28, 29, 87, 139, 198, 206, 266]. On the adaptive landscape we study, this is not the case. Even though larger populations with more clonal

interference experience fewer substitution events, their final fitness is higher. At very high mutation rates, large populations hardly have any substitutions (Figures 3.2d and 3.3d), but they can still achieve a higher final mean fitness (Figures 3.2a and 3.3a). The likely reason is that large populations are more likely to discover beneficial mutations, as long as enough such mutations exist (Figures 3.2b and 3.3b). And when such beneficial alleles occur in a population, they may help increase final mean fitness, even when they do not become fixed. This pattern is consistent with a prevalence of soft selective sweeps [152, p. 472], where multiple beneficial mutations can co-occur and rise in frequency, even though none of them goes to fixation [98, 202].

Second, at large $N\mu$, final mean fitness does not just depend on $N\mu$, but also on population size $N$. Specifically, at a given $N\mu$, larger populations achieve higher mean fitness. The reason is that a high population mutation rate translates into higher mutation rate per individual in smaller populations, which can overwhelm selection.

Third, the mean number of unique sequences explored by an evolving population, as well as the mean final population diversity depend on population size, both for any given $\mu$, and for any given $N\mu$.

Our observations also speak to the question whether adaptive evolution is more rapid in large or small populations, because several conflicting factors can influence the speed of adaptation in such populations [139]. We find that smaller populations have no adaptive advantage over larger populations, because they do not reach higher mean final fitness at any given mutation rate. Thus, even though smaller populations can escape local fitness peaks more easily, have larger neutral neighborhoods (Figures S3.1a and S3.1b), and could thus explore more sequences (Figure 3.4b), they are at a disadvantage, at least on the relatively smooth fitness landscape we study (Figures S3.3 and S3.2).

Theoretical studies that examine the effect of mutation rate and population size on adaptation make assumptions about the prevalence and magnitude of mutational effects. Since ultimately only beneficial mutations increase the average fitness of a population, we focus on models that examine their interactions with mutation rate and population size. Using the terminology introduced by Gillespie [82, 84], we can categorize such theoretical models into the following categories based on the assumptions they make on the prevalence and effect of beneficial mutations: a) strong selection and weak mutation

(SSWM) models, where beneficial mutations are rare and have large fitness effects, b) weak selection and strong mutation (WSSM) models, where beneficial mutations are common, but their fitness effects are small, and c) strong selection and strong mutation (SSSM) models, where beneficial mutations are both common and have large fitness effects. The rate at which beneficial mutations fix in a population depends on the frequency with which they arise and their effect size. A beneficial mutation has initially a probability $2s$ of surviving genetic drift by increasing its frequency to a level at which its fate is only governed by selection [92], where $s$ is the mutation's fitness effect. In an asexual population, an increase in the frequency of a beneficial mutation is not enough for its fixation because it may have to compete with other beneficial mutations for fixation. The probability that two beneficial mutations will co-occur in a population is a function of the population size $N$, the magnitude of their effects $s$, and the beneficial mutation rate $U_b$. Higher $s$ leads to faster fixation of a mutation, and leaves a smaller window for other beneficial mutations to co-occur. Increased $N$ and $U_b$ make it more likely for multiple beneficial mutations to co-occur in a population. Assumptions of these theoretical models affect their predictions. In the SSWM models, beneficial mutations fix one after another in an asexual population [53, 54], because beneficial mutations are rare and it is their mutation rate that limits rate of adaptation. In the WSSM models, many beneficial mutations are present at the same time in a population, and the rate of adaptation is limited by the rate at which these mutations can fix. Accumulating empirical evidence supports high beneficial mutation rates in natural populations [54, 59, 203]. Finally, studies using SSSM models have made different simplifying assumptions that affect their predictions. For example, [81, 263] assumed that beneficial mutations fix separately, but their fitness effects may be different, drawn from an exponential distribution. In contrast, [53, 54] made a model in which all beneficial mutations had the same effect, but could compete with one another for fixation. Both of these models are inconclusive, ignoring the effect of competition among multiple mutations or ignoring the effect of competition among mutations of different effect sizes. Aside from having no a priori assumption about the distribution of fitness effects, our model, using a fitness landscape to simulate population evolution, takes into consideration that the distribution of fitness effects changes as a function of adaptation of a population to an environment. Populations that are far from a fitness peak will experience more and larger effect beneficial mutations than populations already at a fitness peak.

Among the limitations of our study is that we considered only asexual populations. Recombination may alter the evolutionary dynamics substantially [44, 67, 177, 188, 195, 281]. In addition, the landscapes we study are not very rugged, with few fitness peaks for most structures (Figures S3.2 and S3.3), and little reciprocal sign epistasis that might slow down adaptive evolution (Figure S3.3). More rugged landscapes could yield substantially different evolutionary dynamics.

In sum, our observations suggest that simple models of evolutionary dynamics, especially on highly simplified fitness landscapes, need to be taken with caution, because evolutionary adaptation on a complex landscape can reveal interdependencies between various factors affecting adaptive evolution, particularly when $N\mu$ is very large.

## 3.4 Methods

### 3.4.1 Network analysis

We constructed all networks and characterized their graph-theoretical properties using the iGraph library (version 0.7.1) [48] for Python. We used Gephi (version 0.9.1)[16] for network visualization.

### 3.4.2 RNA molecules

Our analysis focuses on two different kinds of RNA molecules. The first kind comprises all RNA molecules of length 10 that have at least a paired base in their minimum free energy (MFE) secondary structure. We chose these short sequences to be able to fully analyze and visualize their genotype space. The second kind comprises a small number of short RNA sequences with biological functions, which we chose from the database of functional RNA molecules fRNAdb [127]. Specifically, we chose four short sequences from different organisms and with different functions, a snmRNA (small non-messenger RNAs), a snoRNA (small nucleolar RNA), a non-coding transcript, and a ribozyme (Table 3.2). The major difference between sequences of length 10 and biological sequences is their length, but this difference may

influence other properties, such as the incidence of neutral and deleterious mutations.

### 3.4.3 Calculating the fitness of RNA sequences

Our measure of fitness is based on the amount of time that an RNA molecule spends in a given structure, such as its minimum free energy (MFE) secondary structure. To calculate the MFE secondary structure of a sequence we used the function `fold` in the ViennaRNA package (version 2.1.9) [150]. To calculate the time that a sequence spends in a given structure (the probability that it is found in this structure at any given time), we used the following procedure. First, we calculated the ensemble free energy *F* of the sequence using again the `fold` program, where *F= -kT ln(Z)* [150]. Here, *Z* is the partition function of the sequence, *k* is the Boltzmann constant ($1.98717 \times 10^{-3}$ *kcal/K*), and *T* is the absolute temperature (310.15 *K* or 37°C in our case) [5]. Thus, the partition function of a sequence is equal to $Z = exp(F/-kT)$. Second, we calculated the free energy *E* of the focal structure using the `energy_of_struct` function within the ViennaRNA package. These calculations also allowed us to compute the probability that the sequence can be found in the focal structure as $p = exp(-E/kT)/Z$ [5]. For a structure whose free energy lies outside an energy interval of $5kT$ (3 $kcal/mol$ at 37°C) above the MFE of the sequence, the time spent in the structure is very small, and we thus set it to zero for the purpose of our simulations.

We used two different measures of fitness, which are both defined relative to an arbitrary target secondary structure *S*. For the first measure, we set an RNA molecule's fitness to zero if its MFE secondary structure was different from *S*. If the molecule's MFE was identical to *S*, we assumed that its fitness was equal to the time that the sequence spent in *S*. We used this measure to calculate the fitness of our RNA sequences of length 10. This measure of fitness ensures that the evolution of RNA populations is confined to the set or network of genotypes that have *S* as their MFE structure.

The second fitness measure, which we used only for the biological sequences, is identical to the first, except that we did not assign sequences whose MFE structure differs from the target structure *S* a fitness value of 0. Instead, we assumed that their fitness is equal to the time they spend in the target structure.

### 3.4.4 Population evolution model

We used only non-modified ribonucleotides [31, 49, 192, 200], i.e. A, C, G and U, in our discrete-time simulations of RNA sequence evolution. Any one evolving population initially consisted of identical sequences whose MFE structure is the target structure for selection. Because we wanted to explore how such sequences evolve towards high fitness, that is, a large fraction of time spent in the MFE structure, we wanted to initialize populations to a state of low fitness. Specifically, in our simulations of sequence evolution for sequences of length 10, we arbitrarily chose a sequence of length 10, whose fitness was in the bottom 5% of the fitness distribution (i.e., it spends little time in its MFE structure) as the initial sequences for each replicate simulation. For each of our 50 replicate evolution simulations of biological sequences, we arbitrarily chose an initial sequence whose fitness was smaller than one percent, i.e. it spent less than 1% of its time in their target structure. The length of this sequence was exactly the same as that of the biological sequence, so that it could in principle fold into the same target structure. Each of these replicate simulations thus started from a different initial sequence, but with otherwise identical parameters.

Our simulations proceeded through repeated cycles ("generations") of mutation and selection. For a given mutation rate $\mu$ per sequence and generation ($0.0001 < \mu < 1$), we mutated individual sequences as follows. We chose a random number $n$ from a Poisson distribution with mean $\mu$ as the number $n$ of nucleotides to be mutated in the sequence. To mutate the sequence, we chose a random position (with a uniform distribution along the sequence) for mutation, replaced its nucleotide by a randomly chosen one of the three possible alternative nucleotides, and repeated this process $n$ times.

After all sequences had been mutated, we determined their fitness, and chose sequences for survival into the next generation by randomly sampling with replacement from the mutated population, where we weighted the probability that a sequence is sampled by its fitness. Sampling with replacement ensures a constant population size across generations.

### 3.4.5 Neutral neighborhood size calculation

We chose 1,000 random sequences and calculated their fitness based on the MFE structure of a reference sequence, which could be one of our natural RNA sequences, or, for sequences of length 10, a sequence with maximum fitness for a given structure. For each of these 1,000 sequences, we calculated the fitness of all one mutant neighbors. If the fitness difference between a sequence and any one of its neighbors was smaller than $1/N$, we considered the neighbor to be in the sequence's neutral neighborhood. We report the average fraction of neighbors of the 1,000 sequences that are neutral.

### 3.4.6 Estimating reciprocal sign epistasis for different sequences

As a measure of landscape ruggedness, we used the fraction of sequences that are separated from their two-mutant neighbors (sequences separated by two single nucleotide changes) by a fitness valley, i.e., where both one-mutant neighbors have lower fitness than the sequence itself and the two-mutant neighbor. As in our other analyses, we considered two fitness values different if they differed by more than $1/N$.

To compute the incidence of reciprocal sign epistasis for any one secondary structure, we first chose from genotype space 1,000 random sequences that were capable of forming this secondary structure. To do so for biological sequences, we generated random RNA sequences (with uniform and independent nucleotide distributions across the nucleotide sites), and verified for each sequence whether it could form the desired structure, until we had identified 1,000 such sequences. We considered a sequence as being able to form the desired structure, if this structure occurred among all structures within an energy interval of $5kT$ above the sequence's MFE structure. For sequences of length 10, we simply chose 1,000 random sequences from each genotype network (or all sequences in the genotype network if the size of the network was less than 1,000). For all 1,000 sequences thus generated, we counted the number of fitness valleys between that sequence and all its two-mutant neighbors that had higher fitness.

### 3.4.7 Computing population diversity

We used the number of sequences that exist in an evolving population in any one generation as a measure of diversity of the population. More specifically, we computed two complementary measures of population diversity. The first is the average number of unique sequences in the last generation (800), where the average is taken over all replicate simulations. The second is the total number of unique sequences that occurred during the entire course of a simulation, i.e., each sequence that existed in a population during at least one generation, averaged over all replicates.

### 3.4.8 Counting the incidence of deleterious, neutral and beneficial mutations

To identify the number and type of mutations that occur in any one generation of a simulation, we tracked every mutation in single sequences that occurred during a simulation. We compared the fitness of a sequence before and after each mutation, and considered the mutation neutral if this difference was less than $1/N$. If the fitness of the sequence increased (decreased) by more than $1/N$ after a mutation, we considered the mutation beneficial (deleterious).

### 3.4.9 Number of substitutions

At each generation of a population's simulation, we considered any mutant sequence as having become fixed if it was different from the founding sequence of the population, and if its population frequency exceeded a value of 90% (following common practice in population simulations [53, 246] to limit computational cost). We counted any sequence fixation event only once. That is, if a sequence exceeded this fixation threshold in any one generation, dropped below this threshold later on, and then exceeded the threshold once again at a later time, we considered that the sequence underwent only one fixation event.

### 3.4.10 Finding network peaks

We used the Python package Genonets[122] to find fitness peaks in the adaptive landscape defined on the genotype network of sequences with the same structure. The package requires a minimal fitness differential $\Delta$ between two neighboring sequences to call two sequences different in their fitness. The smaller this minimal fitness differential, the greater may be the number of apparent peaks in a rugged fitness landscape. We used $\Delta = 0$.

### 3.4.11 Finding the consensus sequence and its distance to the initial sequence

We determined a population's consensus sequence in a given generation in the following way. For every site in the sequence, we identified all alleles present in the population, and counted the absolute frequency of each allele, i.e., the number of individuals that harbored the allele. We assigned the most frequent allele to the consensus sequence at this site. If two or more alleles had the highest absolute frequency, we assigned an 'N' to the consensus sequence at the site. We computed the distance of the consensus sequence to any other sequence as the Hamming distance between the two sequences, i.e., as the number of sites that differed in their nucleotides between them.

## 3.5 Supplementary figures

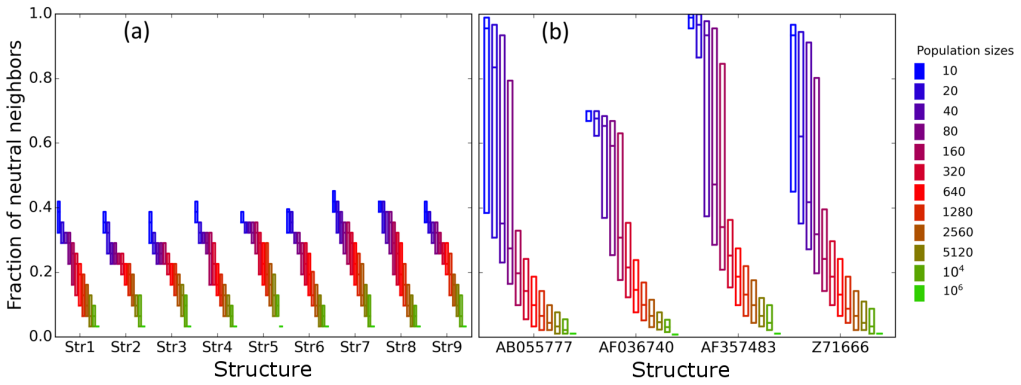


FIGURE S3.1: **Fraction of neutral single-mutation neighbors.** For each of the **(a)** nine secondary structures of length 10 ((Table 3.1) and **(b)** the four biological secondary structures (Table 3.2) (both depicted on horizontal axes), we selected 1,000 random sequences and determined the fraction of neighbors with a fitness difference smaller than $1/N$ for a range of population sizes (legend). In these boxplots, each box encloses the second and third quartiles of the 1000 replicates, and the center line corresponds to the median. As expected, the fraction of neutral neighbors decreases with increasing population size.

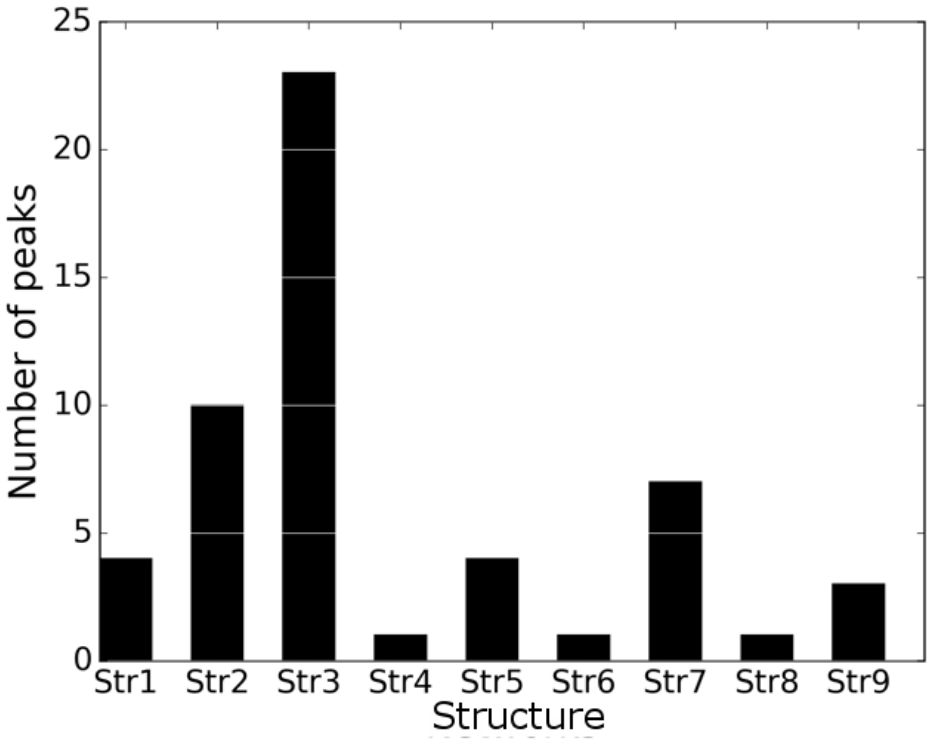


FIGURE S3.2: **Number of fitness peaks for different structures of sequences of length 10.** A peak corresponds to one or more nodes in a fitness landscape, whose neighbors all have lower fitness. (See Methods for peak identification).

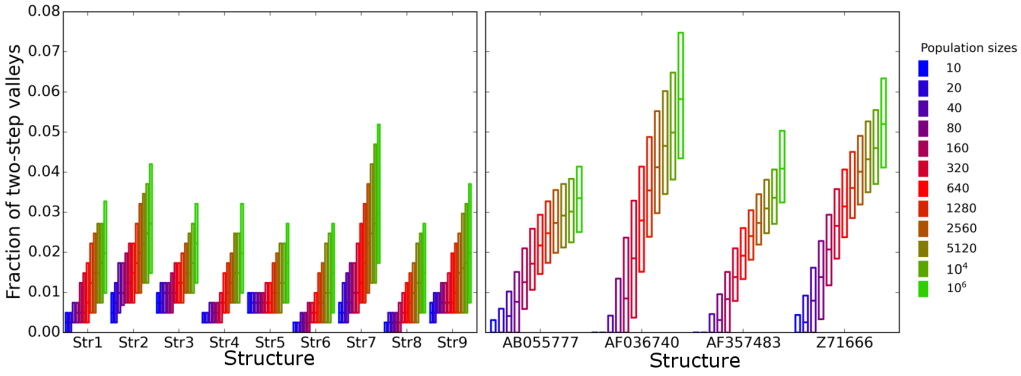


FIGURE S3.3: **Extent of reciprocal sign epistasis in fitness landscapes of biological and length 10 sequences.** The figure shows the estimated fraction of sequence quadruplets with sign epistasis, which is equivalent to the estimated fraction of fitness valleys caused by reciprocal sign epistasis in **(a)** genotype networks of sequences of length 10 (Table 3.1), **(b)** genotype networks of biological sequences (Table 3.2). With increasing population size, the incidence of fitness valleys due to reciprocal sign epistasis increases. However, the overall fraction of such valleys is small (less than 0.06). See methods for the identification of reciprocal sign epistasis.

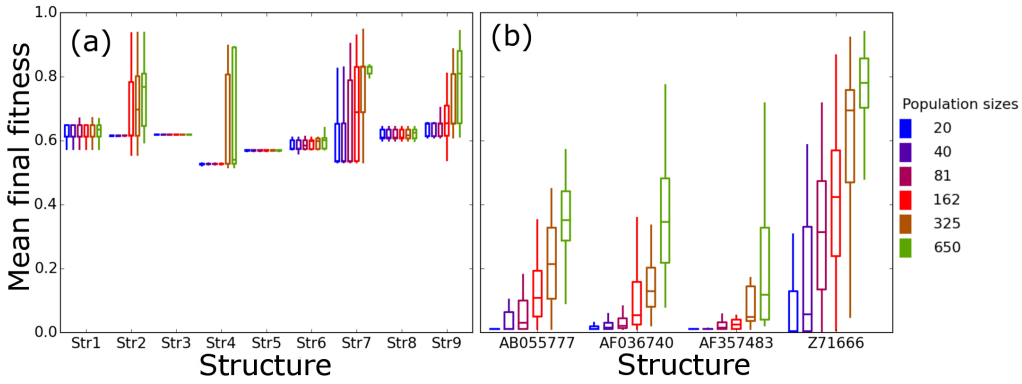


FIGURE S3.4: **Mean population fitness at the end of the simulations at constant $\mu = 0.0001$. (a)** all nine considered RNA structures of length 10 (Table 3.1), **(b)** biological sequences (Table 3.2). We randomly-selected a low-fitness sequence to initialize each simulation, and then simulated 800 generations of mutation and selection. We performed 50 replicate simulations for each population size at a fixed mutation rate of $\mu = 0.0001$ per sequence per generation (see Methods). Each box encloses the second and third quartiles of the 50 replicates, the center line corresponds to the median, and the whiskers depict the minimum and maximum values obtained from any replicate, excluding the outliers.

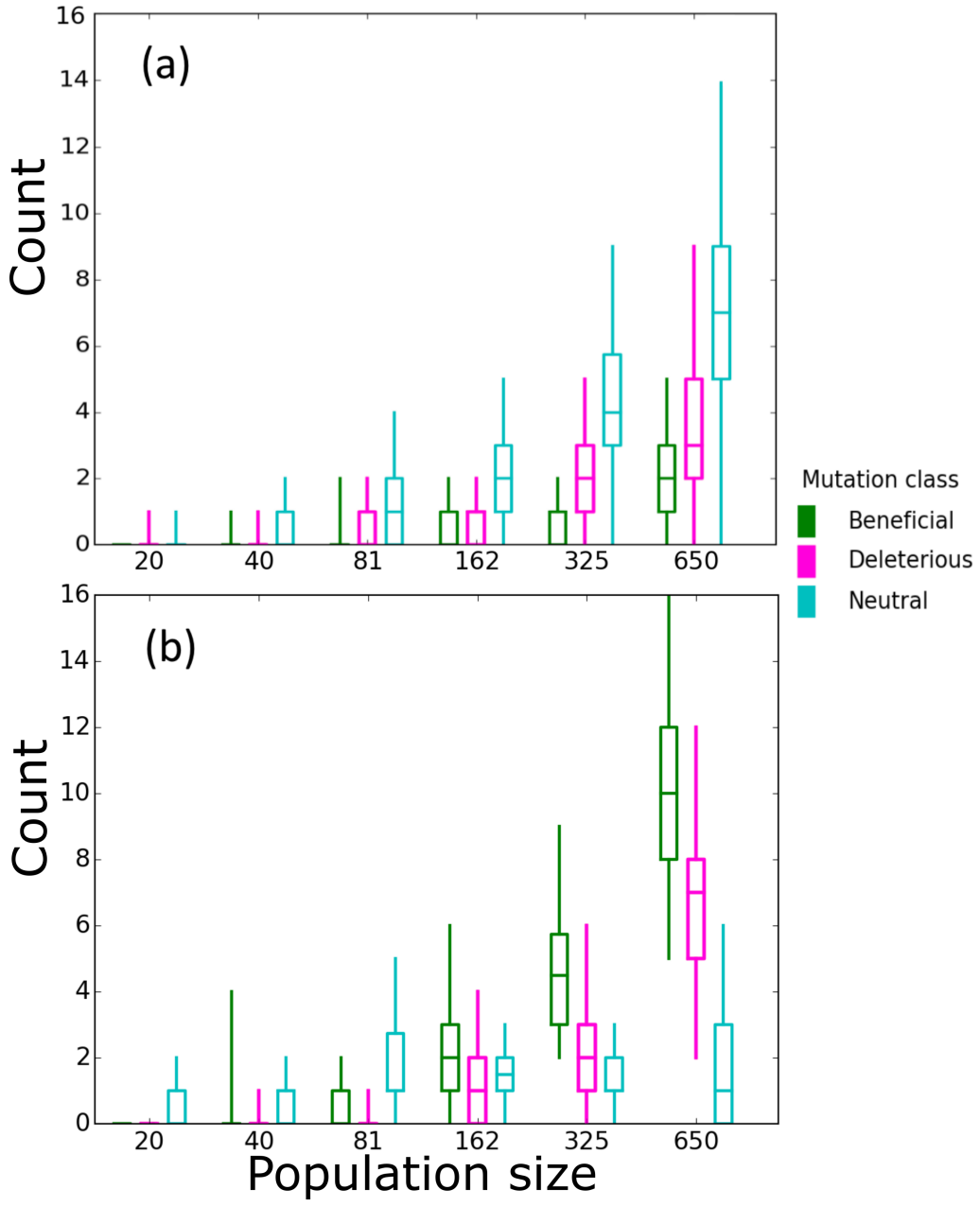


FIGURE S3.5: **Mean numbers of unique beneficial, deleterious, and neutral substitutions for RNA secondary structure 1 (Str1, Table 3.1) and AF036740 (Table 3.2) at $\mu = 0.0001$.** We randomly-selected a low-fitness sequence to initialize each simulation, and then simulated 800 generations of mutation and selection. Fifty replicates were simulated for mutation rate $\mu = 0.0001$ and a range of population sizes (horizontal axes, see Methods). Data are based on the final generation of 50 replicate simulations. In these boxplots, we grouped the data by their fitness effects (beneficial, deleterious, neutral); each box encloses the second and third quartiles of the 50 replicates, the center line corresponds to the median, and the whiskers depict the minimum and maximum values obtained from any replicate, excluding outliers. Interestingly, we observed more deleterious than beneficial mutations in the **(a)** Str1 secondary structure simulations, and more beneficial than deleterious mutations in the **(b)** biological AF036740 secondary structure.

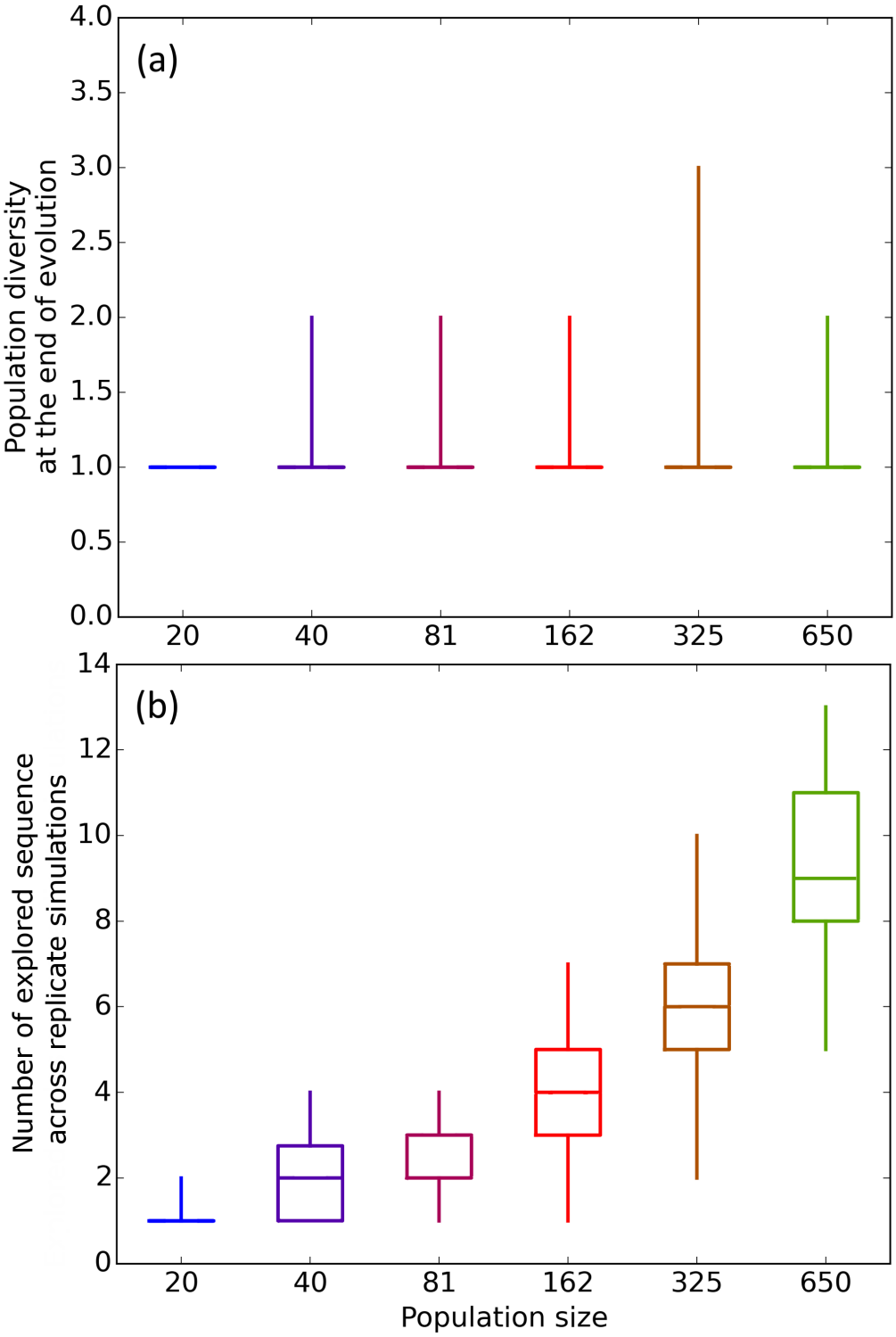


FIGURE S3.6: **Population diversity and sequence exploration in sequences with secondary structure 1 (Str1, Table 3.1) at $\mu = 0.0001$.** We randomly-selected a low-fitness sequence to initialize each simulation, and then simulated 800 generations of mutation and selection. We performed 50 replicate simulations for each population size (horizontal axes, see Methods). Boxplots show **(a)** final population diversity (number of unique sequences at generation 800), and **(b)** total unique sequences explored. Each box encloses the second and third quartiles of the 50 replicates, the center line corresponds to the median, and the whiskers depict the minimum and maximum values obtained from any replicate, excluding the outliers.

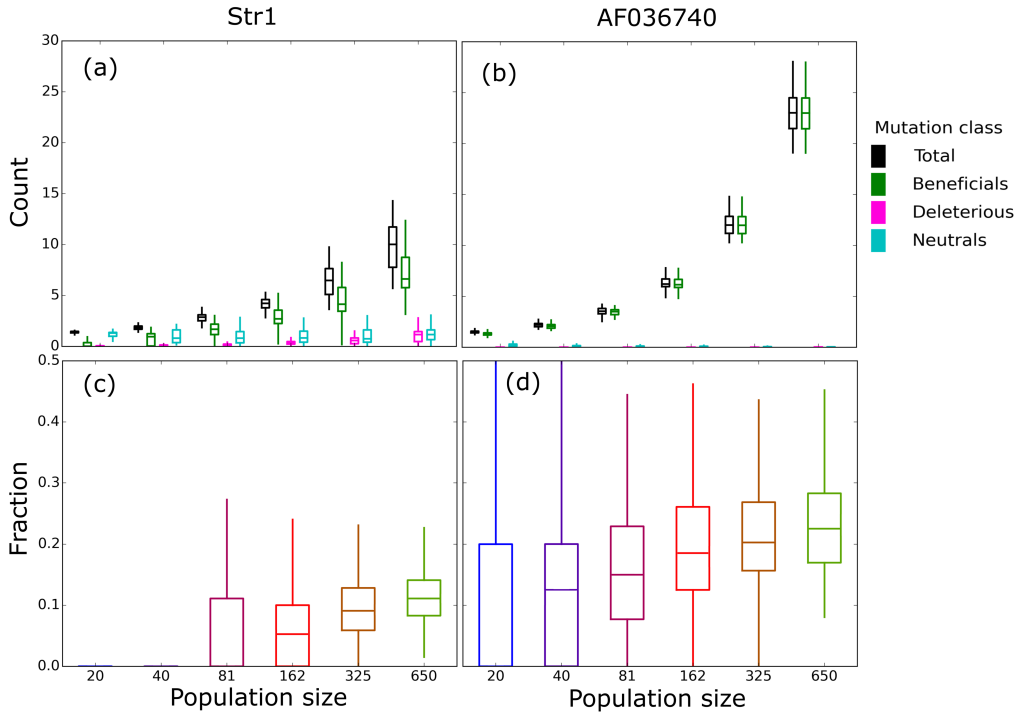


FIGURE S3.7: **Mean numbers of unique beneficial, deleterious, and neutral mutations per generation and fraction of beneficial mutations for RNA secondary structure 1 (Str1, Table 3.1) and AF036740 (Table 3.2) at $\mu = 0.01$.** We randomly-selected a low-fitness sequence to initialize each simulation, and then simulated 800 generations of mutation and selection. Fifty replicates were simulated for mutation rate $\mu = 0.01$ and a range of population sizes (horizontal axes, see Methods). Data are based on the final generation of 50 replicate simulations. Boxplots summarize mean numbers of unique beneficial, deleterious, and neutral mutations for **(a)** Str1 and **(b)** AF036740, and mean fraction of beneficial mutations for **(c)** Str1 and **(d)** AF036740. Each box encloses the second and third quartiles of the 50 replicates, the center line corresponds to the median, and the whiskers depict the minimum and maximum values obtained from any replicate, excluding the outliers.

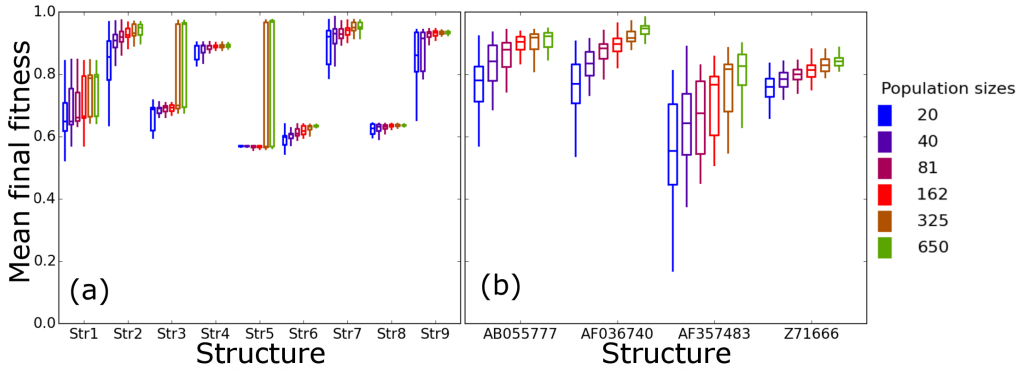


FIGURE S3.8: **Final mean population fitness after evolution of secondary structures of length 10 and (Table 3.1) and biological RNA secondary structures at $\mu = 0.01$.** We randomly-selected a low-fitness sequence to initialize each simulation, and then simulated 800 generations of mutation and selection. We performed 50 replicate simulations for each structure (horizontal axes) and population size (legend, see Methods). Boxplots show mean final population fitness of all the replicates for **(a)** the structures of length 10 and **(b)** the biological RNA molecules. Each box encloses the second and third quartiles of the 50 replicates, the center line corresponds to the median, and the whiskers depict the minimum and maximum values obtained from any replicate, excluding outliers. In all structures, except Str4, the largest population (N=650) has a significantly higher final mean population fitness than the smallest population (N=20) (Mann-Whitney U test, multiple-testing correction according to FDR; Str1: p=$1.01 \times 10^{-7}$; Str2: p=$9.55 \times 10^{-12}$; Str3: p=$8.00 \times 10^{-9}$; Str5: p=$4.81 \times 10^{-3}$; Str6: p=$8.42 \times 10^{-14}$; Str7: p=$1.01 \times 10^{-9}$; Str8: p=$2.00 \times 10^{-2}$; Str9: p=$1.55 \times 10^{-5}$; AB055777: p=$1.16 \times 10^{-13}$; AF036740: p=$1.16 \times 10^{-17}$; AF357483: p=$8.64 \times 10^{-12}$; Z71666: p=$8.03 \times 10^{-16}$).

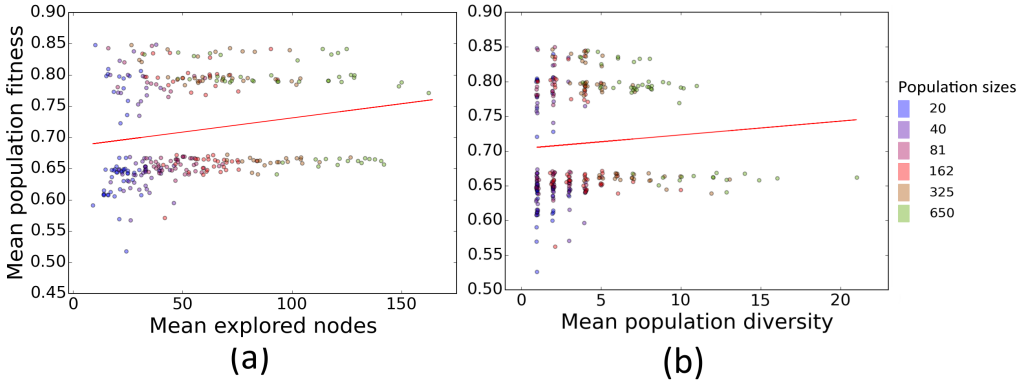


FIGURE S3.9: **Average final fitness is associated with the number of sequences explored.** Results are based on simulations with structure a (Str1, Table 3.1) at constant $\mu = 0.01$ and population size $N$. We randomly-selected a low-fitness sequence to initialize each simulation, and then simulated 800 generations of mutation and selection. Each data point represents the results of one simulation at a given population size (color legend). The vertical axes show mean final population fitness (i.e., fitness at generation 800). To better distinguish data points, a small amount of noise is added to each point. **(a)** Mean final fitness is significantly associated with the total number of unique sequences that the population explored during 800 generations (horizontal axis; Pearson's r=0.19, p=0.00094). **(b)** However, mean final fitness is not significantly associated with the final population diversity, defined as the number of unique sequences at generation 800 (horizontal axis; Pearson's r = 0.079, p = 0.17). The bimodal distribution of mean final fitness evident in both panels is specific to the simulated structure, and not a consistent pattern across different structures.

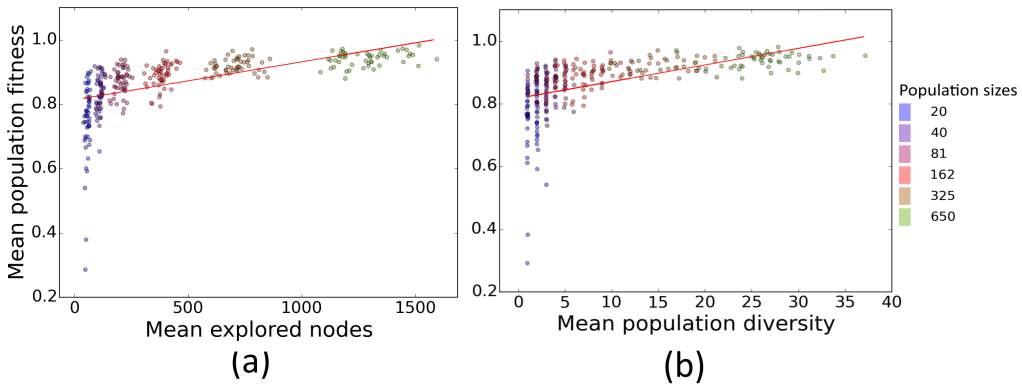


FIGURE S3.10: **Average final fitness is associated with the number of sequences explored and final population diversity for sequences with AF036740 RNA secondary structure.** Results are based on simulated evolution of structure of AF036740 of biological sequences (Table 3.2) at constant $\mu = 0.01$ and population size $N$. We randomly-selected a low-fitness sequence to initialize each simulation, and then simulated 800 generations of mutation and selection. Each data point represents the results of one simulation at a given population size (color legend). The vertical axes show mean final population fitness (i.e., fitness at generation 800). To better distinguish data points, a small amount of noise is added to each point. Mean final fitness is significantly associated with the **(a)** total number of unique sequences that the population explored during 800 generations (horizontal axis; Pearson's r=0.59, p=$4.51 \times 10^{-30}$), and **(b)** the final population diversity, defined as the number of unique sequences at generation 800 (horizontal axis; Pearson's r=0.56, p=$1.21 \times 10^{-26}$).

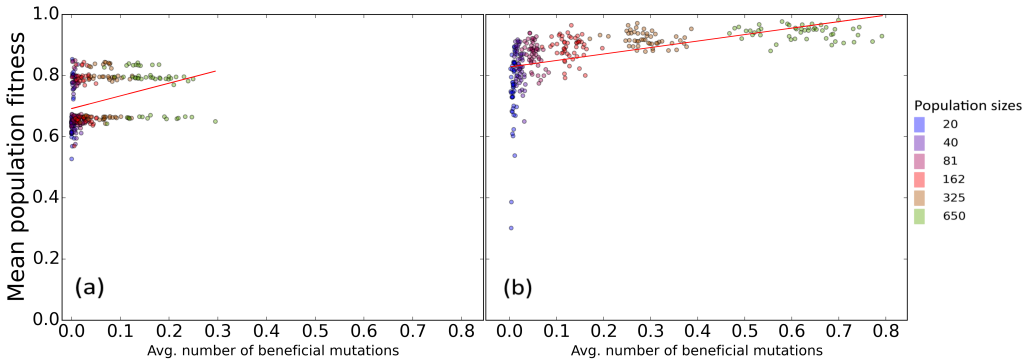


FIGURE S3.11: **Average final fitness is associated with average number of beneficial mutations at $\mu = 0.01$.** Results are based on sequences with **(a)** secondary structure 1 (Str1, Table 3.1) and **(b)** with structure of AF036740 (Table 3.2). Correlations are significant for both samples (Mann-Whitney U test, $p = 8.5 \times 10^{-9}$ and $p = 2.5 \times 10^{-24}$ for (a) and (b), respectively.) We randomly-selected a low-fitness sequence to initialize each simulation, and then simulated 800 generations of mutation and selection. We simulated 50 replicate populations for each structure (horizontal axes) and population size (legend, see Methods). For better presentation of data, small noise is added to each data point. Colors of data points show which population size they represent (color legend).

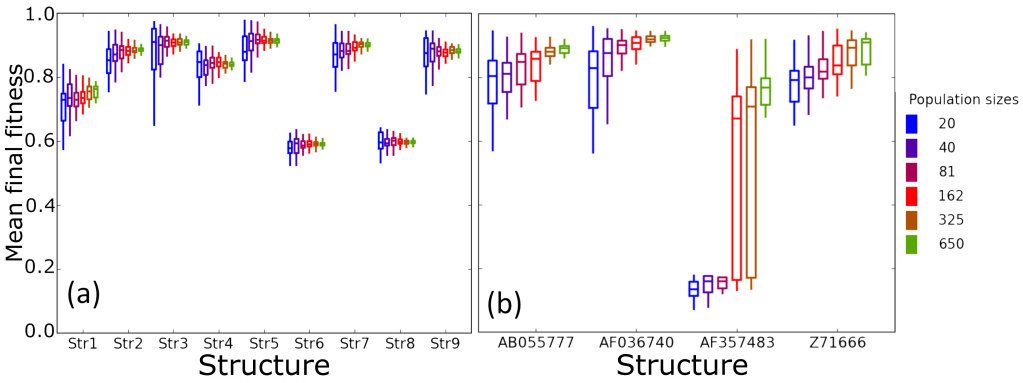


FIGURE S3.12: **Final mean population fitness after evolution of secondary structures of length 10 and (Table 3.1) and biological RNA secondary structures at** $\mu = 0.1$**.** We randomly-selected a low-fitness sequence to initialize each simulation, and then simulated 800 generations of mutation and selection. We performed 50 replicate simulations for each structure (horizontal axes) and population size (legend, see Methods). Boxplots show the final mean population fitness of all the replicates for **(a)** the structures of length 10 and **(b)** the biological RNA molecules. Each box encloses the second and third quartiles of the 50 replicates, the center line corresponds to the median, and the whiskers depict the minimum and maximum values obtained from any replicate, excluding outliers. In all structures, except Str3, Str4, Str5, Str8 and Str9, the largest population (N=650) has a significantly higher final mean population fitness than the smallest population (N=20) (Mann-Whitney U test, multiple-testing correction according to FDR; Str1: p=$4.00 \times 10^{-5}$; Str2: p=$2.22 \times 10^{-4}$; Str6: p=$1.81 \times 10^{-2}$; Str7: p=$6.36 \times 10^{-4}$; AB055777: p=$5.73 \times 10^{-7}$; AF036740: p=$4.97 \times 10^{-12}$; AF357483: p=$4.75 \times 10^{-16}$; Z71666: p=$9.08 \times 10^{-12}$).

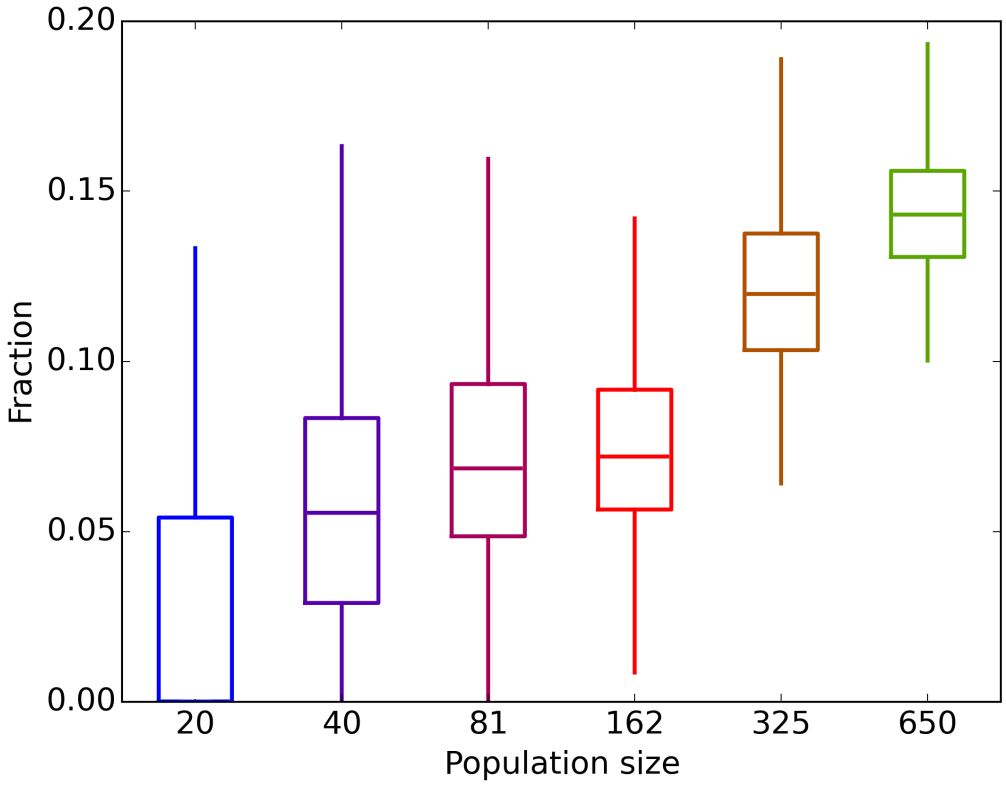


FIGURE S3.13: **Fraction of beneficial mutations for RNA secondary structure 1 (Str1, Table 3.1) at $\mu = 0.1$.** We randomly-selected a low-fitness sequence to initialize each simulation, and then simulated 800 generations of mutation and selection. Fifty replicates were simulated for mutation rate $\mu = 0.1$ and a range of population sizes (horizontal axes, see Methods). Data are based on the final generation of 50 replicate simulations. Boxplots summarize the mean fraction of beneficial mutations. Each box encloses the second and third quartiles of the 50 replicates, the center line corresponds to the median, and the whiskers depict the minimum and maximum values obtained from any replicate, excluding the outliers.

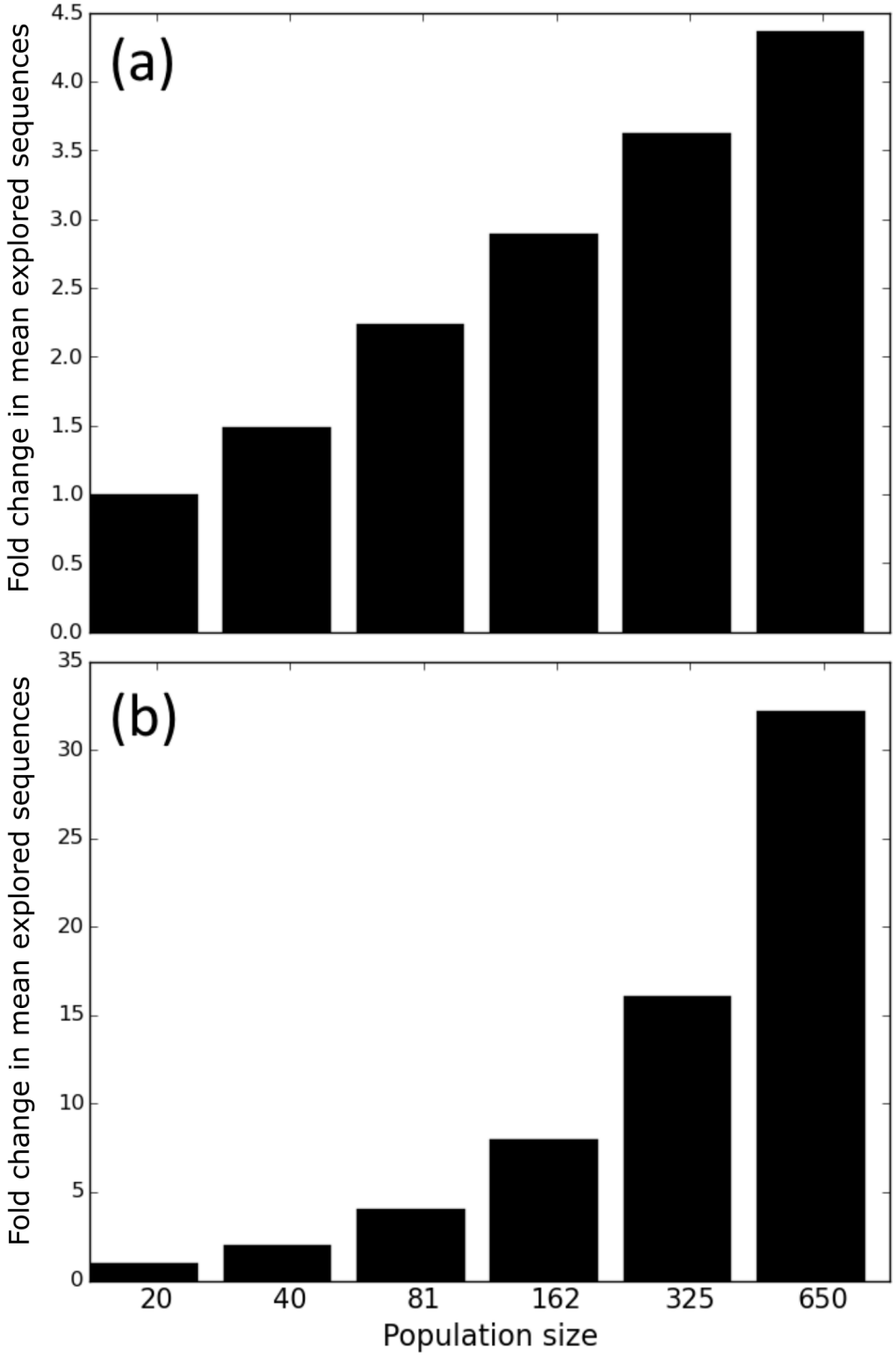


FIGURE S3.14: **Fold change in number of explored sequences for two different structures. (a)** Str1 (Table 3.1), **(b)** AF036740 (Table 3.2). The bars show by how many fold the mean number of explored sequences increases when population size increases, relative to populations of size $N = 20$. The mutation rate is held constant at $\mu = 0.1$ for all populations.

TABLE S3.1: Associations between mean final fitness and population diversity at generation 800 in all 9 structures of length 10 (Table 3.1), when mutation rate $\mu = 0.01$.

| Structure | Pearson's r | p-value |
|---|---|---|
| Str1 | 0.08 | 0.17 |
| Str2 | 0.33 | 2.72e-9 |
| Str3 | 0.53 | 2.63e-23 |
| Str4 | 0.17 | 0.003 |
| Str5 | 0.46 | 9.29e-17 |
| Str6 | 0.49 | 1.36e-19 |
| Str7 | 0.38 | 1.09e-11 |
| Str8 | 0.28 | 4.44e-7 |
| Str9 | 0.30 | 7.49e-8 |

TABLE S3.2: Associations between mean final fitness and explored sequences across generations in all 9 structures of length 10 (Table 3.1), when mutation rate $\mu = 0.01$.

| Network | Pearson's r | p-value |
|---|---|---|
| Str1 | 0.19 | 9.36e-4 |
| Str2 | 0.46 | 7.08e-17 |
| Str3 | 0.69 | 9.38e-44 |
| Str4 | 0.22 | 1.19e-4 |
| Str5 | 0.68 | 1.74e-42 |
| Str6 | 0.65 | 8.42e-37 |
| Str7 | 0.46 | 8.32e-17 |
| Str8 | 0.33 | 3.69e-9 |
| Str9 | 0.41 | 1.42e-13 |

## 3.6 Supplementary tables

# Chapter 4

# Population size affects adaptation in complex ways: simulations on empirical adaptive landscapes

Ali R. Vahdati, Andreas Wagner

# *Abstract*

Do large populations always outcompete smaller ones? Does increasing the mutation rate have a similar effect to increasing the population size, with respect to the adaptation of a population? How important are substitutions in determining the adaptation rate? In this study, we ask how population size and mutation rate interact to affect adaptation on empirical adaptive landscapes. Using such landscapes, we do not need to make many ad hoc assumption about landscape topography, such as about epistatic interactions among mutations or about the distribution of fitness effects. Moreover, we have a better understanding of all the mutations that occur in a population and their effects on the average fitness of the population than we can know in experimental studies. Our results show that the evolutionary dynamics of a population cannot be fully explained by the population mutation rate $N\mu$; even at constant $N\mu$, there can be dramatic differences in the adaptation of populations of different sizes. Moreover, the substitution rate of mutations is not always equivalent to the adaptation rate, because we observed populations adapting to high adaptive peaks without fixing any mutations. Finally, in contrast to some theoretical predictions, even on the most rugged landscapes we study, small population size is never an advantage over larger population size. These result show that complex interactions among multiple factors can affect the evolutionary dynamics of populations, and simple models should be taken with caution.

## 4.1 Introduction

How do mutation rate and population size interact on different landscape topographies to affect a population's adaptation? Answering this question can be important for predicting the evolutionary dynamics of different kinds of populations, such as those of pathogens or endangered species. There are many factors affecting the adaptation of organisms, including the presence or absence of genetic recombination; the structure of the fitness landscape [271], e.g. its shape and size; DNA mutation rates; the distribution of fitness effects of mutations; and effective population size [53, 54, 93, 110, 153, 166, 194, 263]. We focus on two of these factors; namely, effective population size $N_e$ [34, 154] and mutation rate $\mu$, to better understand their role in adaptation on empirical adaptive landscapes. Specifically, we would like to know at which mutation rates and levels of landscape ruggedness smaller or larger populations have an evolutionary advantage. Do smaller populations outcompete larger ones when landscape ruggedness increases? What is the role of mutation rate in the adaptation of populations of different sizes?

Population size has a major impact on evolutionary dynamics. Under some circumstances, it is advantageous for a population to be larger. The reason is that natural selection is more effective in removing weakly deleterious mutations and fixing weakly beneficial mutations [190]. Consequently, the beneficial mutations go to fixation more frequently in larger populations, and deleterious mutations go to fixation less frequently [3, 139]. Additionally, when the product of population size and mutation rate ($N\mu$) is large enough, an evolving population can cross fitness valleys through a process called stochastic tunneling [4, 108, 128, 259, 262]. Specifically, such a population is more likely to produce double mutants that do not experience the deleterious effect of a single mutant, which may allow it to cross a fitness valley [235].

Producing more mutations is not always an advantage. When several beneficial mutations are simultaneously present in an asexual population, they compete with each other for fixation. This slows the time to fixation of a beneficial mutation. This phenomenon is called clonal interference [81], and it can slow down the rate of adaptive substitutions in a population [36]. Producing fewer mutations per generation, smaller populations are less likely to be affected by clonal interference, and they may thus adapt faster [81, 235]. Furthermore, genetic drift is stronger in smaller populations. In a rugged

landscape, where achieving a higher fitness likely requires passing through fitness valleys, strong genetic drift facilitates valley crossing [93, 110]. Moreover, some fitness valleys for large populations become flat for smaller populations, because any fitness difference between two mutations smaller than $1/N$ becomes invisible to selection [110, 137, 190, 235].

The many factors affecting evolutionary dynamics often interact in non-intuitive ways to define the evolutionary outcome of a population. Therefore, most previous theoretical studies include simplifying assumptions to model the role of one or a few of these factors [29, 53, 54, 123, 137, 153]. Examples include epistatic interactions among mutations [45, 248], and the distribution of fitness effects [46, 68, 239], which define the ruggedness of a fitness landscape. For example, [93] used randomly generated fitness landscapes to study the effect of population size on the evolution of microbes; and [110] used a three-locus model with arbitrary fitness values for each genotype to study the advantage of small populations on rugged landscapes. Another example is an assumed distribution of fitness effects with rare beneficial mutations to predict the association between the substitution rate of beneficial mutations and the population size [139]. Whether beneficial mutations are rare depends on the proximity of a population to a fitness peak. Violation of such assumptions can lead to dramatically different evolutionary outcomes [139]. In experimental studies, where realistically complex fitness landscapes are examined [135, 216], researchers have inevitably limited knowledge about, and control over, underlying evolutionary mechanisms, such as the distribution of fitness effects and the mutational trajectories of a population. This is because such fitness landscapes are usually large, and the possibilities to replicate experiments and to vary parameters are limited.

For these reasons, some studies make contradictory observations about the effect of population size on adaptation. For example, the rate of adaptation, defined as the number of beneficial substitutions, has been predicted to increases with effective population size $N_e$ [139]. However, this prediction only holds when beneficial mutations are rare. The frequency of beneficial mutations, in turn, depends on the location of a population on a fitness landscape and on the topology of the landscape [139]. Thus, some studies have found associations between the $N_e$ and rate of adaptation [55], while others have not [9, 80, 117]. Our study tries to fill the gap between theoretical and experimental studies, using a system where we have more knowledge about, and

control over, important factors such as population mutation rates, evolutionary trajectories, and the identity of substituted genotypes, than experimental systems. At the same time, we need to make fewer ad hoc assumptions than most previous theoretical studies. One of these assumptions is the distribution of fitness effects. In an empirical landscape, this distribution changes as a population approaches a fitness peak. For example, when a population gets closer to a peak, beneficial mutations become rarer, without the need to make ad hoc assumptions about their frequency.

We consider 957 empirical adaptive landscapes [2]. Each landscape encompasses the binding affinity of a transcription factor to all of its cognate DNA sequences (i.e., binding sites). These binding affinities are derived from protein binding microarrays in the form of an enrichment score (E-score), which describes the relative binding preference of a transcription factor to all possible DNA sequences of length eight [20]. The topographies of these landscapes have recently been characterized in rich detail [2], which provides an opportunity to study how the topographies of empirical adaptive landscapes interact with $N$ and $\mu$ to affect the adaptation rate of an evolving population. Transcription factor binding affinity is an important molecular phenotype, because it can affect gene expression. For example, increasing the affinity of an activating transcription factor's binding site will decrease the factor's disassociation rate, thereby increasing the rate of transcription of the downstream gene. If increased expression is selectively advantageous in a given environment (e.g., an antibiotic resistance gene in the presence of an antibiotic), then increased binding affinity may confer increased fitness. The importance of high binding affinity transcription factor binding sites is evidenced by their signature of positive selection in microbes and humans [178, 179], as well by their proximity to actively transcribed genes in the embryo of *Drosophila melanogaster* [144]. We therefore use binding affinity as a proxy for fitness.

Using these empirical adaptive landscapes, we do not make many ad hoc assumptions about the distributions of fitness effects, the structure of the landscape, or epistatic interaction among mutations, because such information is implicitly present in the landscapes. We simulate populations with a range of mutation rates $\mu$ and population mutation rates $N\mu$, and analyze all mutational trajectories of populations during their evolution. We find that mutation rate $\mu$ and population mutation rate $N\mu$ are not always sufficient parameters to predict the adaptation rate of populations on these landscapes.

Population diversity and the extent of landscape exploration, rather than the substitution rate of mutations, can affect the adaptation rate.

## 4.2 Results

### 4.2.1 Structure of binding affinity landscapes

From the 1,137 landscapes studied in [2], we simulated the evolution of populations on those 957 landscapes that had at least 100 sequences. We then chose nine of these landscapes for a more detailed analysis. The nine landscapes differ in their ruggedness, as measured by their number of peaks. A peak is defined as a set of sequences whose affinity is larger than that of all their neighboring sequences [122]. Table 4.1 lists the names of these nine transcription factors, their DNA binding domains, the species they belong to, and their number of peaks.

TABLE 4.1: Landscapes in our study. Each column describes the following information: 'TF name': name of the transcription factor to which the sequences bind; 'Species': the species in which the transcription factor occurs; 'Number of components': number of connected components within each network, i.e., components in which sequences are accessible from one another through a path of one or more links; 'Network size': total number of sequences in landscape; 'Size of the dominant genotype network': number of sequences in the largest connected component; 'Number of peaks': number of peaks in the landscape (see Methods); 'Study': the study from which data were retrieved for constructing the landscape.

| **TF name** | **Species** | **Number of components** | **Network size** | **Size of the dominant genotype network** | **Number of peaks** | **Study** |
|---|---|---|---|---|---|---|
| NCU03110 | *Neurospora crassa* | 1 | 1,064 | 1,064 | 1 | [260] |
| TIFY2B | *Arabidopsis thaliana* | 1 | 1,050 | 1,050 | 1 | [260] |
| NCU06990 | *Neurospora crassa* | 1 | 1,038 | 1,038 | 2 | [260] |
| AZF2 | *Arabidopsis thaliana* | 1 | 1,051 | 1,051 | 3 | [260] |
| Six6 | *Mus musculus* | 3 | 658 | 656 | 6 | [10] |
| NCU00445 | *Neurospora crassa* | 4 | 589 | 586 | 7 | [260] |
| KDM2B | *Homo sapiens* | 6 | 634 | 629 | 9 | [260] |
| FBXL19 | *Tetraodon nigroviridis* | 7 | 730 | 724 | 13 | [260] |
| kdm2aa | *Danio rerio* | 13 | 513 | 499 | 36 | [260] |

Some landscapes have multiple connected components, i.e. sets of nodes (sequences) that are reachable from one another through a sequence of single step mutations. We call the largest of these components the dominant component and limit our simulations to these dominant components. The single step mutations we consider are either point mutations, or single base pair insertions / deletions [2, 201]. The landscapes comprise between 513 and 1,064 sequences, and have between 1 and 13 connected components (Table 4.1). Figure 4.1 shows one of the landscapes used in this study, that of the *Arabidopsis thaliana*'s transcriptional repressor AZF2. Each circle represents a sequence and links connect sequences that differ by a single mutation.

The evolutionary dynamics of a population on an adaptive landscape depends in part on the average fraction of neutral neighbors of its genotypes. When genotypes in a population have larger neutral neighborhoods, the population may be able to explore a larger fraction of the landscape without facing deleterious mutations. Hence, it may more easily discover beneficial mutations and new phenotypes [5]. Neutral neighborhood size is a function of effective population size $N_e$ [96], which equals consensus population size $N$ in our simulations, because our simulated populations experience no population size fluctuations. We analyzed the size of each neutral neighborhood in different landscapes and with different population sizes. We consider the fitness difference of any two neighboring sequences neutral if it is smaller than $1/N$ [124, 191]. Figure S4.1 shows the fraction of neutral neighbors among all nodes in a landscape, for all nine different landscapes and different population sizes. As expected, neutral neighborhood size decreases with increasing population size, which makes it more difficult for larger populations to evolve neutrally and cross fitness valleys [5].

We used a variation of the Wright–Fisher model (see Methods) to evolve populations on our landscapes for 1,000 generations of mutation and selection, which favors increases in binding affinity. We performed 100 replications for each simulation. Since we are interested in analyzing the effect of population size $N$ and mutation rate $\mu$ on the adaptation of populations, we systematically explored a range of mutation rates ($0.001 \leq \mu \leq 1$) and population mutation rates ($0.01 < N\mu < 10$) with seven population sizes ($10 < N < 640$). We chose a maximum population size of 640 based on the size of the landscapes, so that even in a high mutation regime, only a fraction of the landscape would be occupied by a population.

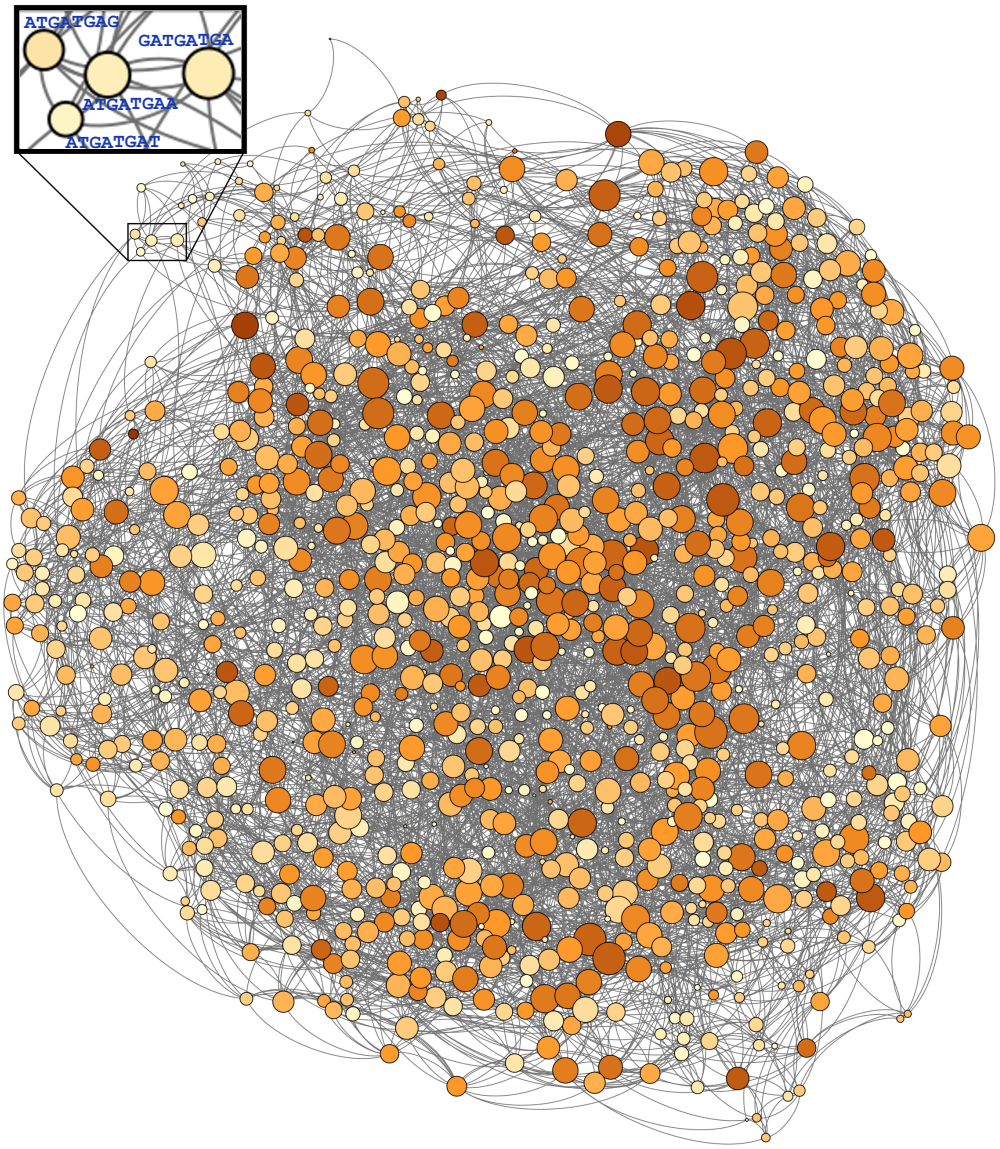


FIGURE 4.1: **The adaptive landscape of the AZF2 transcription factor.** Each node corresponds to a DNA sequence. Two nodes are connected if they differ by a single point mutation or a single indel. Node color corresponds to the affinity of the sequence (Darker=Higher), and node size corresponds to the number of neighbors of the node (Bigger=More). The inset shows that two nodes are connected if they differ by a single mutation. Our display allows for overlapping nodes, so the actual number of nodes may be greater than the number of nodes that are visible.

### 4.2.2 Landscape ruggedness strongly affects adaptation

We initially determined whether the measurement of ruggedness in these landscapes, namely the number of peaks, affects evolutionary dynamics. To that end, we simulated evolution on all of the 957 landscapes [2]. We analyzed correlations between the mean final affinity of simulated populations, normalized by the maximum binding affinity in each landscape, and the number of peaks in each landscape, and at different mutation rates. In line with our expectation, populations in more rugged landscapes have lower mean population affinity at the end of simulations (i.e. generation 1,000) (Table S4.1). In more rugged landscapes, populations are more likely to get trapped on local optima, and this may be a bigger problem for larger populations, because drift is weaker for them compared to smaller populations. These observations hold for all mutation rates ($\mu = 0.001$ - $\mu = 1$).

We also asked whether the size of (number of sequences in) the global peak of each landscape correlates with the mean final affinity of the populations. We found strong and positive correlations (Table S4.2): the larger the size of the global peak of a landscape, the higher the mean final affinity of a population. This indicates that larger peaks are easier to find.

### 4.2.3 Adaptive evolution under varying mutation rate $\mu$

We first investigated how interactions between different mutation rates $\mu$ and population sizes $N$ affect population adaptation, using a range of mutation rates between $\mu = 0.001$ and $\mu = 1$.

$\mu = 0.001$

At this low mutation rate, the population mutation rate is $N\mu << 1$ for all population sizes. Larger populations consistently achieve higher mean binding affinity at the end of simulated evolution (Figure 4.2a). Larger populations have several advantages to help them find adaptive peaks better than smaller populations, even at mutation rates this small. First, since larger populations have a higher population mutation rate $N\mu$, they are slightly more diverse at any generation (Figure 4.2b). Second, and consequently, larger populations visit more unique sequences (Figure 4.2c). They are therefore better at exploring the landscape, which gives them more opportunities

for identifying adaptive peaks. Third, and in line with the second observation, larger populations fix more mutations, most of which are beneficial (Figure S4.2). This is because they experience more mutations, and because selection is more effective in larger populations [110, 137, 235].

$\mu = 0.01$

At a mutation rate of $\mu = 0.01$, we still find that larger populations have higher mean binding affinity at the end of the evolutionary simulations than smaller populations, although the difference between larger populations is smaller than at $\mu = 0.001$ (Figures 4.3a and S4.3). At this mutation rate, populations fall into two evolutionary regimes. Specifically, for four population sizes ($N = 10$, $N = 20$, $N = 40$, and $N = 80$) $N\mu < 1$, and for the other three ($N = 160$, $N = 320$, and $N = 640$) $N\mu > 1$. When there is more than one lineage harboring a beneficial mutation, these lineages compete with each other for fixation, resulting in slower fixation rates of either lineage, a phenomenon called clonal interference [81]. When $N\mu > 1$, populations are polymorphic most of the time, which increases the likelihood of clonal interference [198]. We first tested whether we find clonal interference in these populations, and if it increases with population size. Figure 4.4 shows the average number of unique mutations that are simultaneously present in the population, and the effect of these mutations, i.e. beneficial, deleterious or neutral, relative to the ancestral sequence of the population. The average number of unique mutations at each generation, and the average number of beneficial unique mutations, increases with population size. Consistent with the existence of clonal interference, we find that the number of beneficial substitutions for most landscapes (all except FBXL19 and kdm2aa) is an increasing function of $N$ when $N\mu < 1$ ($N = 10$, $N = 20$, $N = 40$, and $N = 80$), but a decreasing function of $N$ when $N\mu > 1$ ($N = 160$, $N = 320$, and $N = 640$) (Figure S4.4). Moreover, despite fixing no more or even fewer beneficial mutations than smaller populations due to increased clonal interference, larger populations reach higher mean final binding affinity. To explain this pattern, we pooled data from all simulations, and asked whether the mean final population binding affinity correlates with two measures of population diversity, i.e. the number of explored sequences during the evolutionary simulation and the amount of standing variation at the final generation. We found strong positive associations between both metrics of diversity and mean final binding affinity (Tables S4.4 and S4.5). Note that larger populations are both more diverse

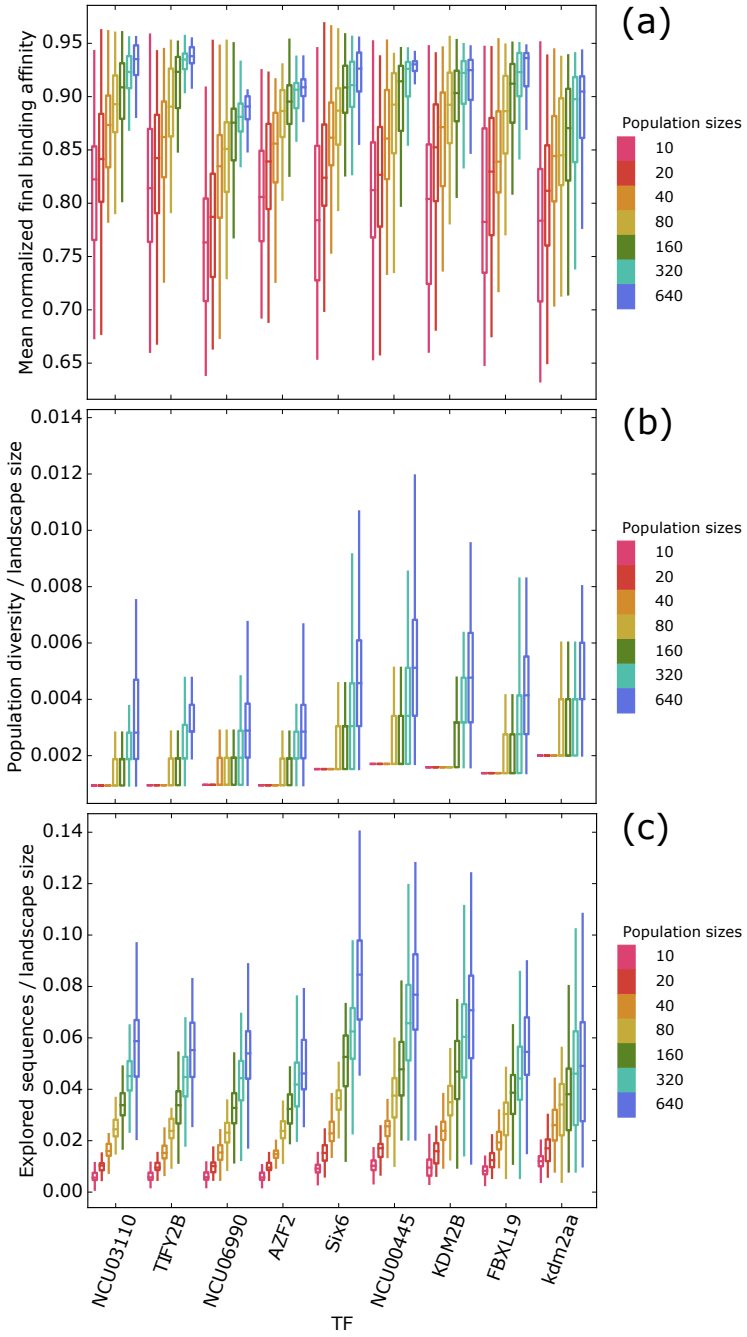


FIGURE 4.2: **Mean final binding affinity, sequence exploration and diversity of populations at $\mu = 0.001$.** The figure shows **(a)** the mean population binding affinity at the end of the simulations, **(b)** the population diversity at the end of the simulations, i.e. the number of unique sequences at generation 1,000, and **(c)** the total number of unique sequences visited by a population during 1,000 generations. Data in (a) are normalized by the maximum affinity value in each landscape, data in (b) and (c) are normalized by landscape size. Horizontal axes on all panels show different transcription factor affinity landscapes ordered from left to right in increasing order of ruggedness. We randomly-selected a sequence of low binding affinity to initialize each simulation, and then simulated 1,000 generations of mutation and selection. We performed 100 replicate simulations for each population size at a fixed mutation rate of $\mu = 0.001$ per sequence per generation (see Methods). Each box encloses the second and third quartiles of data from 100 replicates, the center line corresponds to the median, and the whiskers depict the minimum and maximum values obtained from any replicate, excluding outliers.

in the last generation (Figure 4.3b) and explore more sequences during evolution (Figure 4.3c). These observations suggest that, unsurprisingly, larger populations have more standing variation, which increases the prevalence of beneficial mutations (Figure S4.5), which in turn is strongly associated with increased mean population binding affinity (Table S4.6). In sum, the mean final binding affinity of evolving populations is not completely determined by the number of beneficial substitutions, but also by the population diversity.

$\mu = 0.1$

At a mutation rate of $\mu = 0.1$, the population mutation rate is $N\mu > 1$ for all populations, and clonal interference is prevalent in all populations, but becomes stronger in larger populations (Figure S4.6). The largest populations ($N = 160$, $N = 320$, and $N = 640$), therefore, have nearly no substitutions (Figure S4.7). Still, they arrive at a higher mean binding affinity than smaller populations (Figure 4.5a). The largest populations in some landscapes (FBXL19, NCU00445, and TIFY2B), however, do not differ in their mean final binding affinity.

Population diversity can help explain how larger populations reach higher mean binding affinity levels, despite fixing nearly no mutations. Larger populations explore more sequences than smaller populations, and the difference in this exploration ability between larger and smaller populations is greater at $\mu = 0.1$ (Figure 4.5c). Similarly, the difference between the fraction of beneficial mutations among all mutations that occur in larger populations and in smaller populations is greater at $\mu = 0.1$ (compare Figures S4.5 and S4.8).

$\mu = 1$

At this large mutation rate, where on average every sequence mutates in every generation ($N\mu >> 1$), we do not find striking differences between the mean final binding affinity at different population sizes (Figure 4.6a). Only the two smallest populations ($N = 10$ and $N = 20$) have a slightly lower mean binding affinity than larger populations. More pronounced, however, is a drop in mean final binding affinity of all population sizes compared with $\mu = 0.1$ (compare Figure 4.5a with 4.6a). This is because of the high fraction of mutant individuals that are created generation. When a population finds and moves to a sequence with a high binding affinity, it will not stay there,

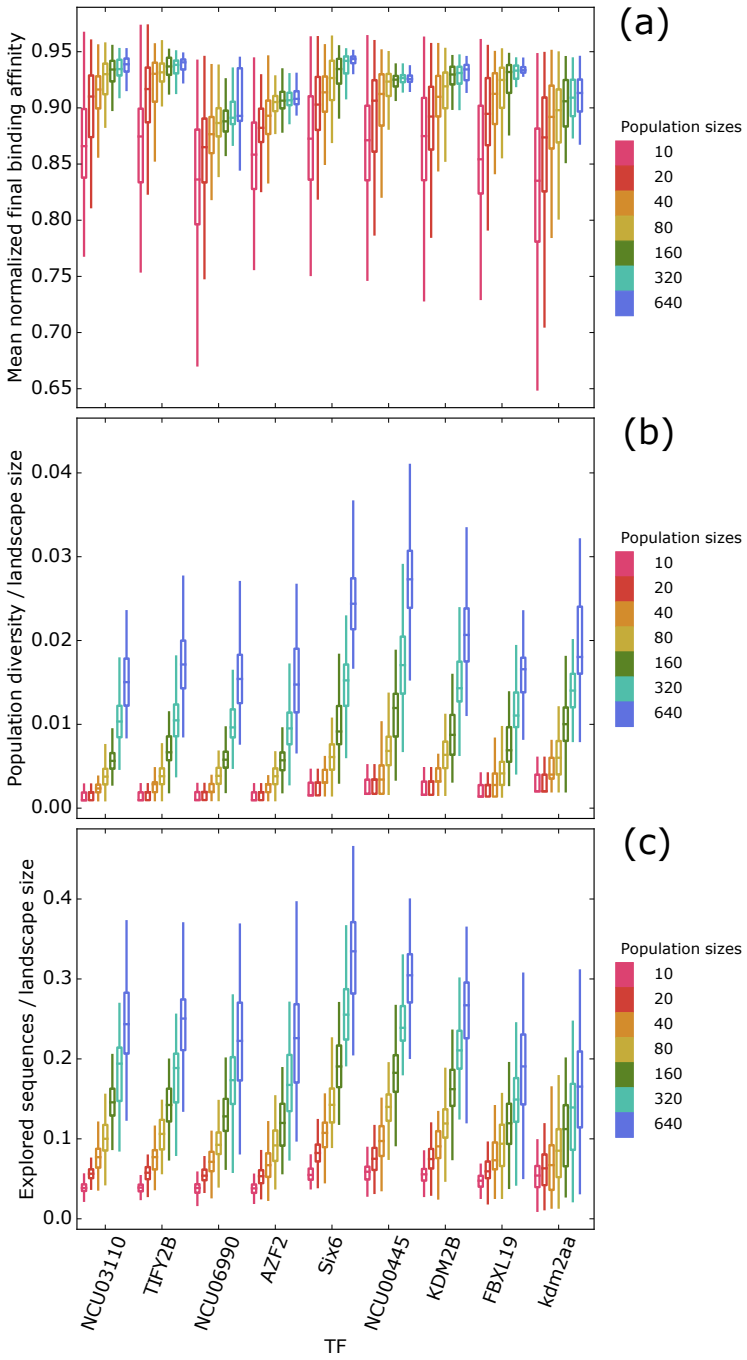


FIGURE 4.3: **Mean final binding affinity, sequence exploration and diversity of populations at $\mu = 0.01$.** The figure shows **(a)** the mean population binding affinity at the end of the simulations, **(b)** the population diversity at the end of the simulations, i.e. the number of unique sequences at generation 1,000, and **(c)** the total number of unique sequences visited by a population during 1,000 generations. Data in (a) are normalized by the maximum affinity value in each landscape, data in (b) and (c) are normalized by landscape size. Horizontal axes on all panels show different transcription factor affinity landscapes ordered from left to right in increasing order of ruggedness. Each box encloses the second and third quartiles of data from 100 replicates, the center line corresponds to the median, and the whiskers depict the minimum and maximum values obtained from any replicate, excluding outliers. Population evolution was simulated in the same way as explained in the caption of Figure 4.2, except that $\mu = 0.01$.

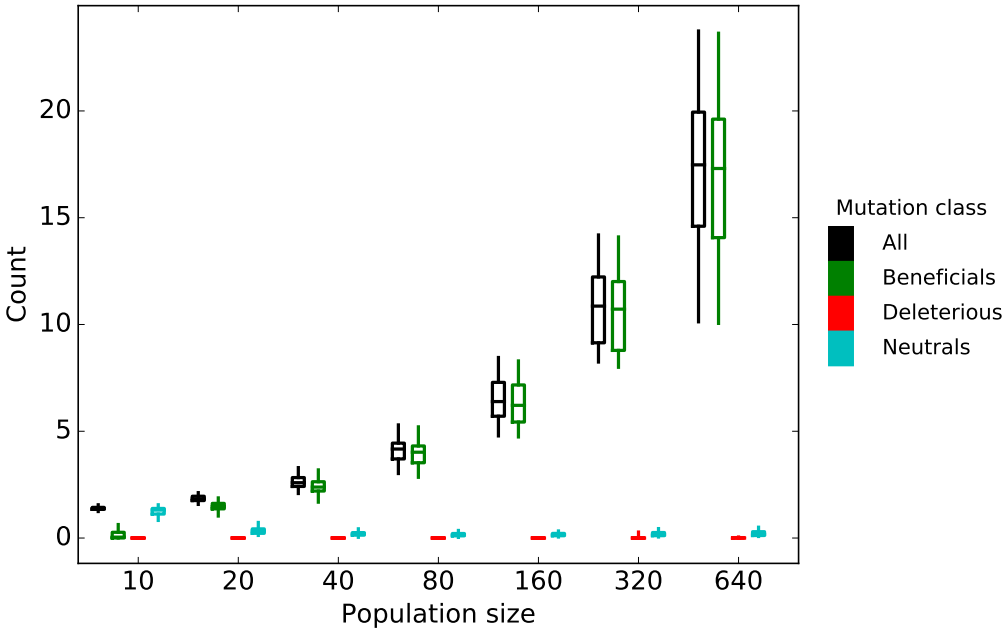


FIGURE 4.4: **More beneficial mutations coexist in larger populations evolving on the AZF2 landscape at constant $\mu = 0.01$.** Boxplots summarize mean numbers of unique total, beneficial, deleterious, and neutral mutations that coexist per generation (color legend) for populations of different sizes (horizontal axis) evolved on the AZF2 landscape. When more than one beneficial mutation is present at the same time in a population, those mutations compete for fixation (clonal interference), resulting in longer fixation time for the mutation that finally fixes in the population. We determined the effect of each mutation compared to the ancestral sequence starting the population simulation. Effects smaller than $1/N$ are neutral. Each box encloses the second and third quartiles of data from 100 replicates, the center line corresponds to the median, and the whiskers depict the minimum and maximum values obtained from any replicate, excluding outliers. Population evolution was simulated in the same way as explained in the caption of Figure 4.2, except that $\mu = 0.01$.

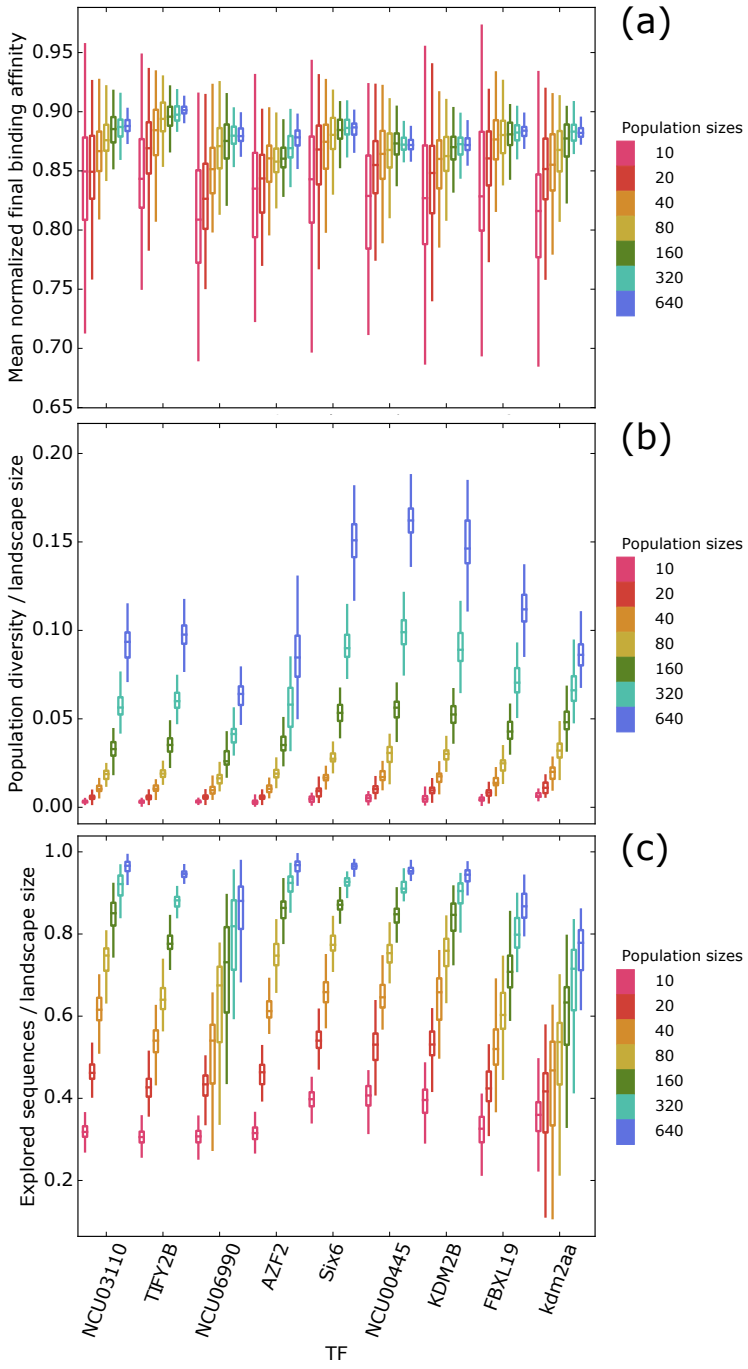


FIGURE 4.5: **Mean final binding affinity, sequence exploration and diversity of populations at $\mu = 0.1$.** The figure shows **(a)** the mean population binding affinity at the end of the simulations, **(b)** the population diversity at the end of the simulations, i.e. the number of unique sequences at generation 1,000, and **(c)** the total number of unique sequences visited by a population during 1,000 generations. Data in (a) are normalized by the maximum affinity value in each landscape, data in (b) and (c) are normalized by landscape size. Horizontal axes on all panels show different transcription factor affinity landscapes ordered from left to right in increasing order of ruggedness. Each box encloses the second and third quartiles of data from 100 replicates, the center line corresponds to the median, and the whiskers depict the minimum and maximum values obtained from any replicate, excluding outliers. Population evolution was simulated in the same way as explained in the caption of Figure 4.2, except that $\mu = 0.1$.

because at the next generation, most individuals mutate away from it. Therefore, the mean affinity of populations fluctuates around lower values and the highest possible mean affinities cannot be attained.

### 4.2.4 Adaptive evolution under varying population mutation rates $N\mu$

Another important quantity in population genetics is the population mutation rate $N\mu$. In the following sections, we will analyze the effect of $N\mu$ on adaptive evolution to find out whether it alone can explain the difference in adaptation between populations of different sizes.

$N\mu = 0.01$ and $N\mu = 0.1$

At these low population mutation rates, populations of all sizes reach similar mean final binding affinity levels (Figures 4.7 and 4.8). Likewise, the extent of sequence exploration (Figures S4.9 S4.10) and population diversity in the last generation (Figures S4.11 and S4.12) is similar among populations of all different sizes. This suggests that $N\mu$ may be adequate to explain evolutionary dynamics when $N\mu$ is not too large.

$N\mu = 1$ and $N\mu = 10$

At the moderate population mutation rate of $N\mu = 1$, we find that the smallest populations (i.e. $N = 10$, $N = 20$, and $N = 40$) are not reaching the same mean final binding affinity as larger populations (Figure 4.9). At the high population mutation rate $N\mu = 10$, this dependency of final fitness on pouplation size is even stronger (Figure 4.10). In addition, there is a negative association between sequence exploration and population size (Figures S4.13 and S4.14). This is likely due to larger neutral neighborhood that is characteristic of smaller populations (Figure S4.1). Larger neutral neighborhoods mean that more neutral mutations are available to smaller populations (Figures S4.15 and S4.16), which thus face fewer limitations exploring novel sequences. Such larger neutral neighborhoods also result in more neutral substitutions in smaller populations (Figures S4.17 and S4.18). Larger populations experience (Figure S4.19) and fix more beneficial mutations than

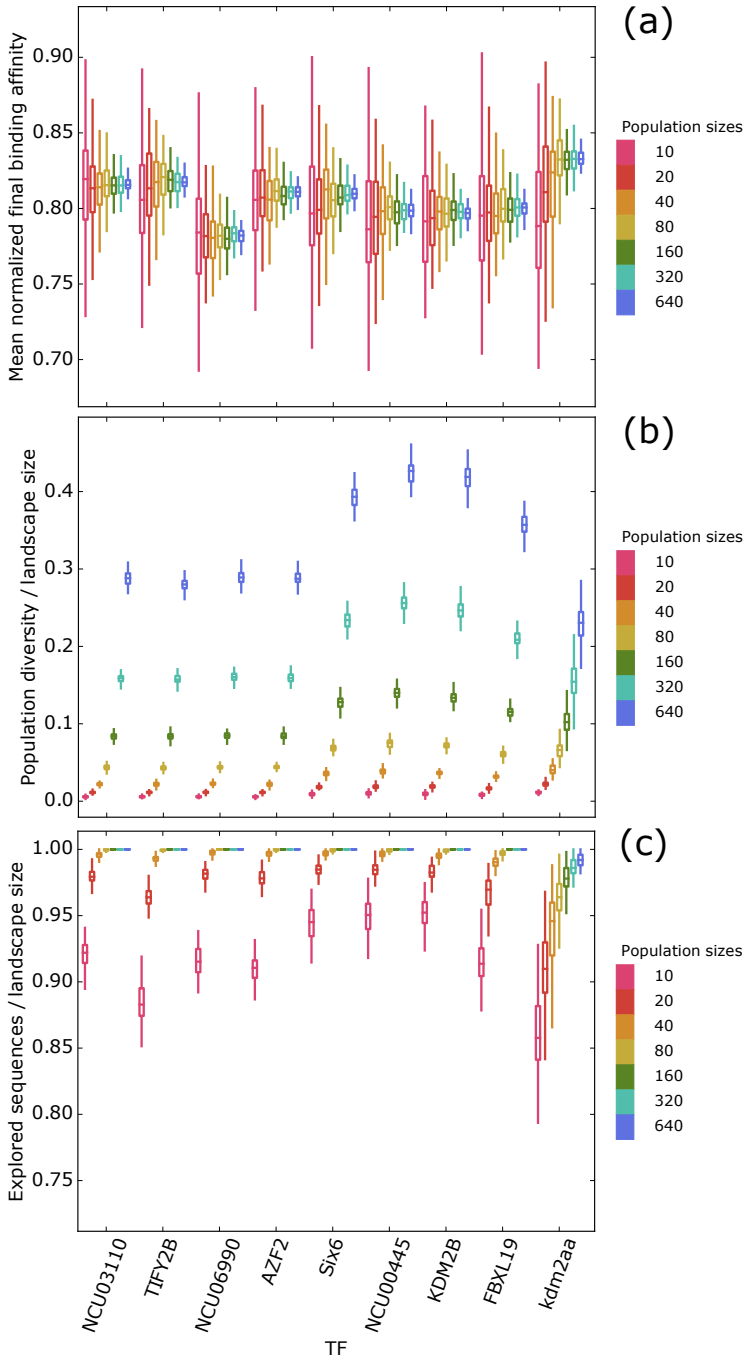


FIGURE 4.6: **Mean final binding affinity, sequence exploration and diversity of populations at $\mu = 1$.** The figure shows **(a)** the mean population binding affinity at the end of the simulations, **(b)** the population diversity at the end of the simulations, i.e. the number of unique sequences at generation 1,000, and **(c)** the total number of unique sequences visited by a population during 1,000 generations. Data in (a) are normalized by the maximum affinity value in each landscape, data in (b) and (c) are normalized by landscape size. Horizontal axes on all panels show different transcription factor affinity landscapes ordered from left to right in increasing order of ruggedness. Each box encloses the second and third quartiles of data from 100 replicates, the center line corresponds to the median, and the whiskers depict the minimum and maximum values obtained from any replicate, excluding outliers. Population evolution was simulated in the same way as explained in the caption of Figure 4.2, except that $\mu = 1$.

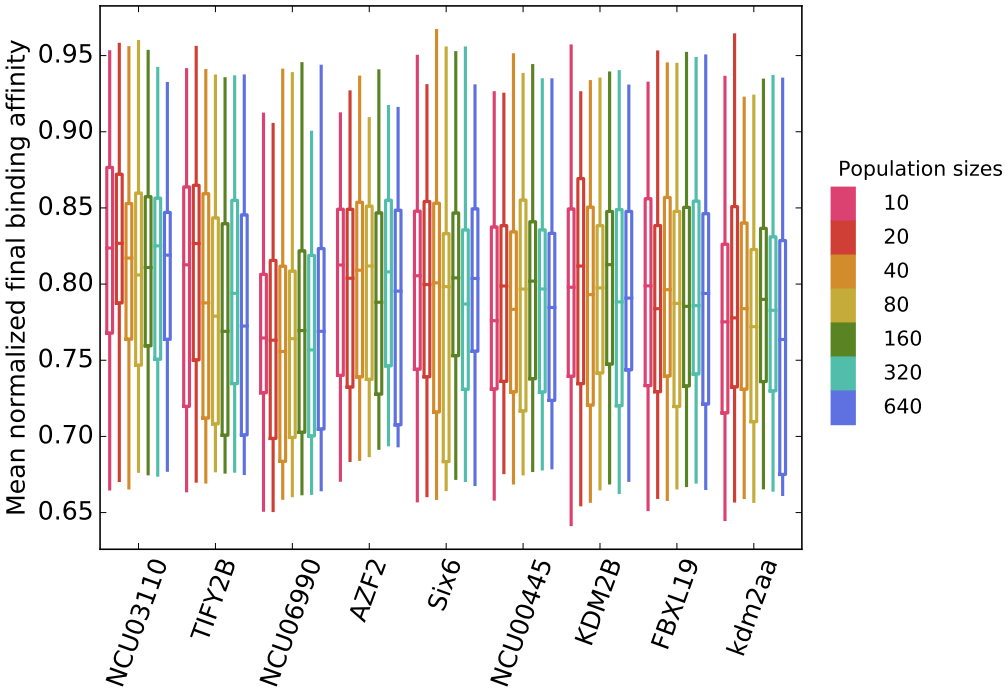


FIGURE 4.7: **Mean population binding affinity at the end of the simulations at constant $N\mu = 0.01$.** We randomly-selected a sequence of low binding affinity to initialize each simulation, and then simulated 1,000 generations of mutation and selection. We performed 100 replicate simulations for each population size at a fixed population mutation rate of $N\mu = 0.1$ per sequence per generation (see Methods). Each box encloses the second and third quartiles of data from 100 replicates, the center line corresponds to the median, and the whiskers depict the minimum and maximum values obtained from any replicate, excluding outliers. Data are normalized by the maximum binding affinity in the landscape.

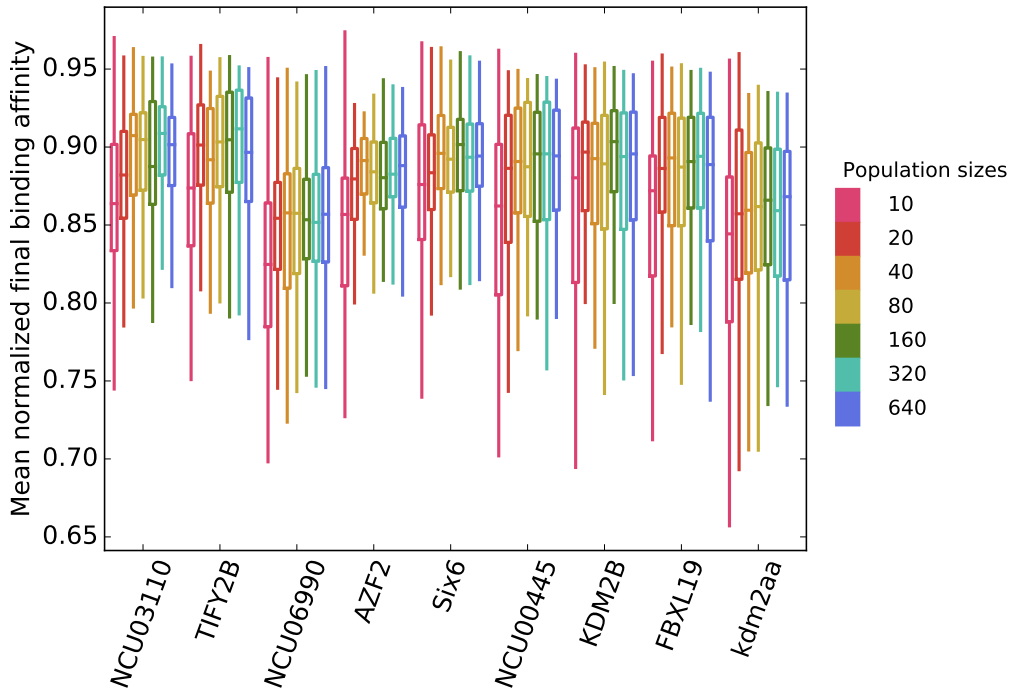


FIGURE 4.8: **Mean population binding affinity at the end of the simulations at constant $N\mu = 0.1$.** Each box encloses the second and third quartiles of data from 100 replicates, the center line corresponds to the median, and the whiskers depict the minimum and maximum values obtained from any replicate, excluding outliers. Population evolution was simulated in the same way as explained in the caption of Figure 4.7, except that $N\mu = 0.1$. Data are normalized by the maximum binding affinity in the landscape.

smaller populations (Figure S4.20) at $N\mu = 1$. At $N\mu = 10$, however, we observe a peak in the maximum fraction of beneficial mutations that the populations experience at intermediate population sizes (Figure S4.21). All populations at $N\mu = 10$ fix fewer mutations than at $N\mu = 1$, but larger populations fix more beneficial mutations (Figure S4.22). Two factors can explain the difference in mean final binding affinity between smaller and larger populations at constant and large population mutation rates. First, selection is more effective at fixing beneficial mutation in larger population. Second, and more importantly, the constant high population mutation rate has a negative effect on the ability to reach high mean affinity for smaller populations, but not for larger populations. A value of $N\mu = 10$ means that an average of ten new mutations are introduced into a population each generation. For a population of size 10, this means that at every generation all individuals are mutated. In a population of size 20, half of all individuals are mutated, but in a population of size 640, only a fraction of 0.016 of individuals are mutated. The high number of mutations overwhelms selection in small populations, making it difficult for small populations to follow a gradual affinity–increasing path.

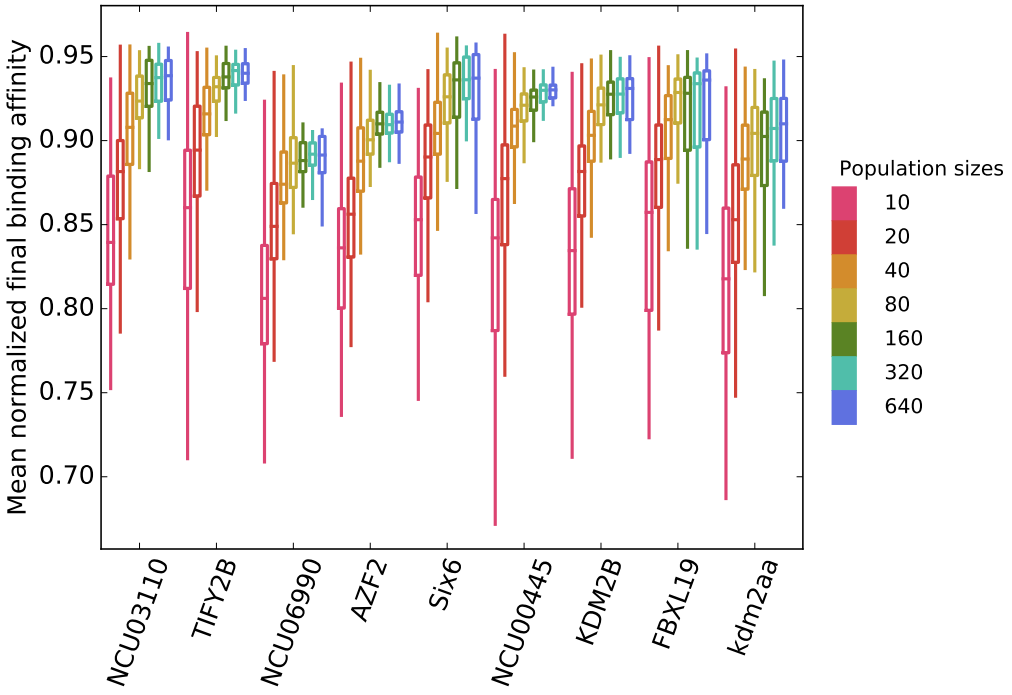


FIGURE 4.9: **Mean population binding affinity at the end of the simulations at constant $N\mu = 1$.** Each box encloses the second and third quartiles of data from 100 replicates, the center line corresponds to the median, and the whiskers depict the minimum and maximum values obtained from any replicate, excluding outliers. Population evolution was simulated in the same way as explained in the caption of Figure 4.7, except that $N\mu = 1$. Data are normalized by maximum binding affinity in the landscape.

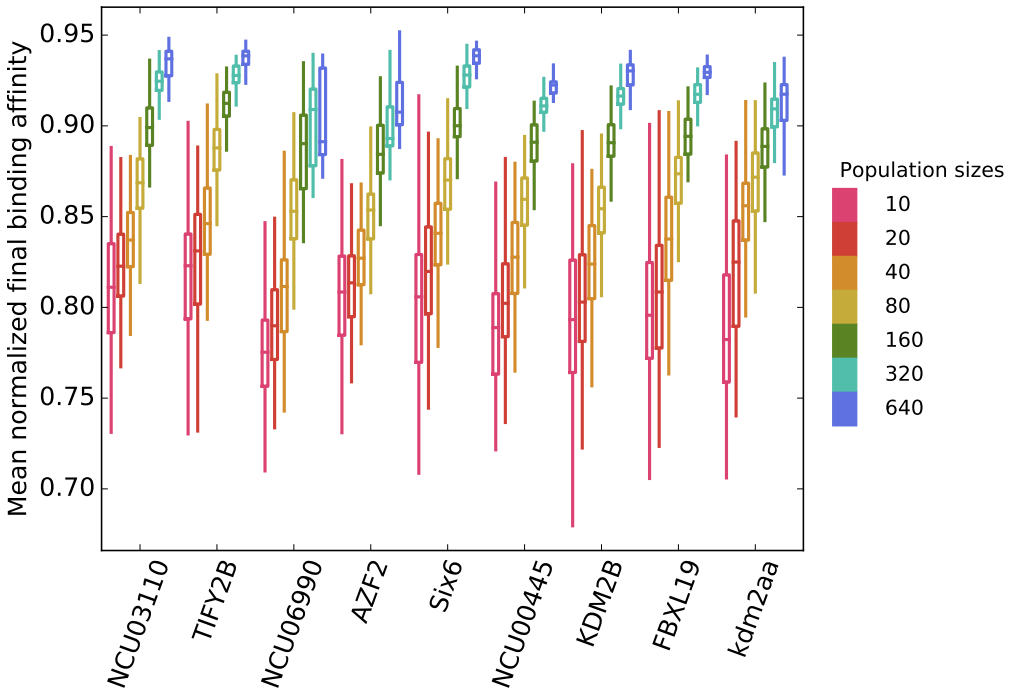


FIGURE 4.10: **Mean population binding affinity at the end of the simulations at constant $N\mu = 10$.** Each box encloses the second and third quartiles of data from 100 replicates, the center line corresponds to the median, and the whiskers depict the minimum and maximum values obtained from any replicate, excluding outliers. Population evolution was simulated in the same way as explained in the caption of Figure 4.7, except that $N\mu = 10$. Data are normalized by maximum binding affinity in the landscape.

## 4.3 Discussion

To understand the rate and limitations of organismal adaptation is a central to evolutionary biology [14, 76, 157, 194, 234, 247, 264]. Efforts to increase our understanding in this area can be divided into two major classes based on their methodology. The first uses theoretical approaches [29, 53, 54, 153]. Due to the complex interactions between different factors, such as mutation rate, changes in effective population size, recombination rate, etc., these approaches usually make many simplifying assumptions, which may not always hold in biological populations. The second class uses experiments [64, 137, 141, 142], which can examine a biological system in its full complexity. However, they provide limited knowledge about the important evolutionary mechanisms, such as the effects of mutations on a population's trajectories, and a fitness landscape's structure. In addition, the ability to replicate experiments and to test different parameters in them is limited.

Here, we used a system that bridges these two approaches. We simulated evolving populations on 957 empirical adaptive landscapes of transcription factor binding sites, and analyzed the evolutionary dynamics on nine such landscapes [2]. We considered the binding affinity between transcription factor and DNA sequences as a proxy for fitness. With such landscapes, we did not have to make ad hoc assumptions about epistatic interactions between mutations, about the distribution of fitness effects, or about landscapes structures. Additionally, we could study the effects of all mutations, and could examine the and mutational trajectories of populations in detail. We found complex interactions between mutation rate and population size, as described below.

Firstly, we found that at any mutation rate, larger populations are better at increasing their mean final affinity (Figure 4.5a). This is intriguing, because at high $N\mu$, due to increased clonal interference, large populations hardly fix any mutations (Figure S4.7); and because the substitution rate, especially that of beneficial mutations, is commonly treated as a measure of adaptation rate [28, 29, 87, 139, 198, 206, 266]. The likely reason that substitution rate does not always determine adaptation is this: Larger populations are more diverse at any given time, and thus explore more sequences in a landscape than smaller populations, which means that they can find beneficial mutations more easily. The presence of multiple beneficial mutations in a population helps the population increase its mean binding affinity, even if no mutation is fixed.

This is akin to a soft selective sweep [152, p. 472], where multiple beneficial mutations occur and increase their frequency in a population without any of them being fixed [98, 202].

Second, we found that even at constant $N\mu$ and for different population sizes, when $N\mu$ is large enough, smaller populations fail to find adaptive peaks as effectively as larger populations (Figure 4.10). The reason is that at constant population mutation rates, smaller populations have a higher mutation rate per genotype than the larger populations. This higher mutation rate overwhelms the small populations and prevents them from following an affinity-increasing path.

Third, we found that sequence exploration and population diversity almost always depend on population size $N$, even when population mutation rates $N\mu$ are constant (Figure S4.14). The only exception is when the population mutation rate $N\mu$ is so low that all populations explore equally few sequences (Figure S4.9).

In sum, we found that smaller populations have no adaptive advantage over larger ones, even when $N\mu$ is constant for populations at different sizes, because smaller populations do not have higher mean final affinity at the end of our simulations. This observation holds regardless of landscape ruggedness, because the landscapes we studied varied in their ruggedness (Table 4.1). In theory, smaller populations could have several advantages on rugged landscapes [216], such as higher chances of escaping local optima, and larger neutral neighborhoods, which could help them explore more sequences, some of which could boost their adaptation. However, these advantages did not lead to better adaptation on the landscapes studied here.

Previous theoretical models on the effect of population size and mutation rate on adaptation make different simplifying assumptions regarding the effect size of beneficial mutations and their prevalence. Such simplifying assumptions change how the models predict evolutionary dynamics. In our study, using empirically motivated fitness landscapes, we did not make such assumptions. Furthermore, the distribution of fitness effects in our model changes as a population ascends a fitness peak. Gillespie introduced the following terminology to differentiate between models adaptive evolution based on their assumptions about the distribution of fitness effects: models that assume strong selection and weak mutation (SSWM), models that assume weak selection and strong mutation (WSSM), and models that assume

both strong selection and strong mutation (SSSM). SSWM models assume that beneficial mutations are so rare that they occur and fix one at a time. These mutations all go to fixation because their effect is large [53, 54]. On the other extreme of the spectrum, WSSM models assume that beneficial mutations are very common, and that many beneficial mutations coexist in a population, at any one time competing for fixation. Consequently, multiple beneficial mutations can go to fixation simultaneously. There is recent experimental evidence showing that beneficial mutations can be more common than previously though [54, 59, 203]. Studies using SSSM models differ in their predictions based on specific assumptions. Some studies assume that the magnitude of the effect of beneficial mutations follows an exponential distribution, but the mutations fix one at a time [81, 263]. Others allow multiple beneficial mutations to be present in a population, but they assign a fixed effect to all such mutations [53, 54]. The validity of this fixed fitness effect for all beneficial mutations depends on the distribution of fitness effects. The two models are, however, not mutually exclusive, and they likely explain only part of the evolutionary dynamics in natural populations.

Our study has limitations, which can be alleviated in future work. Firstly, we studied clonal populations with no recombination. It would be interesting to see how populations adapt on our landscapes in the presence of recombination, because recombination can dramatically affect evolutionary dynamics [44, 67, 177, 188, 195, 281]. Moreover, we used the number of peaks as a measure of landscape ruggedness. It would be interesting to compare the topology of these landscapes with random landscapes used in previous studies, where smaller populations do have an adaptive advantage over larger ones. For example, [93] constructed random landscapes with different numbers of peaks (ruggedness). They simulated populations evolving on the landscapes, and observed that on landscapes with a minimum amount of ruggedness, smaller populations can reach a higher final fitness, because they do not get trapped on local peaks. The conditions that give advantage to smaller populations in such theoretical studies may also exist in other empirical landscapes. Furthermore, we assumed a linear relationship between the binding affinity of a transcription factor to its binding sites and fitness. Although there are examples of increased fitness due to increased binding affinity, the exact form of the affinity-fitness relation is not known. Considering a nonlinear relationship between binding affinity and fitness can change landscape structure, which could affect our observations. It would also be interesting to

analyze the effect of different nonlinear relationships between binding affinity and fitness on landscape structure.

In sum, our results show that in empirical adaptive landscapes, there are complex interdependencies between population size and mutation rate that affect evolutionary dynamics, especially at high $N\mu$, suggesting that conclusions from simplified models should be taken with caution.

## 4.4 Methods

### 4.4.1 Genotype network construction and analysis

Genotype networks were constructed as described in [2, 201]. The data for these networks come from in vitro studies that assess the binding affinity of a transcription factor [140] to all possible DNA sequences of length 8 using protein binding microarrays [19, 20]. The total genotype space consists of 32,896 sequences ($(4^8-4^4)/2+4^4$), where the factor $1/2$ accounts for the merging of sequences with their reverse complement. The number $4^4$ accounts for palindromic sequences, which are identical to their reverse complement and therefore cannot be merged [2]. Reference [2] constructed and analyzed 1,137 binding affinity landscapes from 129 different eukaryotic species and 62 DNA binding domain structural classes. For each transcription factor, a protein binding microarray measures the binding affinity of all 8-mers to the factor. The affinity is represented as a rank-based enrichment score (E-score), which is a variant of the Wilcoxon-Mann-Whitney statistic [20]. This E-score ranges between -0.5 (lowest affinity) to 0.5 (highest affinity). We use the E-score as a proxy for binding affinity, and consider only sequences whose E-score is above 0.35 bound by a transcription factor [2]. We use this threshold because it has yielded a false discovery rate below 0.001 in 104 mouse transcription factors [11]. After identifying a set of sequences that bind each transcription factor, we constructed genotype networks for each transcription factor. The nodes of the network are DNA sequences. Two nodes are connected by a link if they differ by a single mutation. The single mutations considered are either point mutations or single nucleotide insertions / deletions. We characterized graph-theoretical properties of these networks using the iGraph library (version 0.7.1) [48] for Python. We used Gephi (version 0.9.1)[16] for network visualization.

### 4.4.2 Population evolution model

Each landscape only includes sequenes bound by a single transcription factor. However, the total number of sequences of length 8 used in the study (32,896 sequences, either bound to a transcription factor or not bound to any of factors), comprises a bigger network, which we call the network of all possible mutations. For simulations on each landscape, we initialized evolving populations with sequences of low binding affinity, because we wanted to explore the dynamics of populations evolving towards high binding affinity. Specifically, we started each simulation by choosing an arbitrary sequence from the bottom 5% of sequences, according to their E-scores, as the starting sequence of the simulation. Our simulations are limited to the dominant component within each landscape. We initialized a population of $N$ individuals with the same initial sequence. For each set of parameters, we performed 100 simulation replicates, and for each replicate we simulated 1,000 generations of mutation and selection. At each generation, we determined how many mutations each sequence would experience by drawing from a Poisson distribution with a mean equal to the mutation rate $\mu$ of the population. If a sequence was to experience one mutation, we chose randomly one of its neighbors in the landscape. If it was to experience two mutations, we first randomly chose one of its neighbors, and then randomly chose one of the neighbors of the neighbor as the mutant, excluding the original sequence (thus prohibiting back mutations), and likewise for any additional mutations. After the mutation step, we assigned a value $l$ to each sequence by choosing a random number from a uniform distribution in the range of the sequence's E-score$\pm\Delta$. $\Delta$ is a parameter specific to each landscape, which defines a threshold to call two E-scores different in a protein binding microarray experiment, E-scores of each sequence are measured by two replicates, and $\Delta$ is the residual standard error of the linear regression between the E-scores of all bound sequences in the two replicate measurements [2]. Finally, as the selection step, we randomly sampled exactly $N$ sequences from all the sequences with replacement, where the probability of sampling each sequence was weighted by its value of $l$. We note that with this selection method, population sizes remain constant every generation.

### 4.4.3 Neutral neighborhood size calculation

For each landscape, we considered the binding affinity of all neighbors of each of a landscape's sequences. If the binding affinity difference between the sequence and its neighbor was smaller than $1/N$, the neighbor is part of the neutral neighborhood of the sequence. We report the fraction of neutral neighbors of all sequences in each landscape.

### 4.4.4 Computing population diversity

We computed two measures of population diversity. The first measure corresponds to the number of unique sequences at the last generation in each simulation. We report its average across 100 simulation replicates. The second measure is the total number of unique sequences that were visited by a population across all generations, averaged over 100 simulation replicates.

### 4.4.5 Counting the incidence of deleterious, neutral, and beneficial mutations

To calculate the incidence of deleterious, neutral, and beneficial mutations in each population, we tracked every mutation. If the binding affinity difference of sequence and its mutant (whose affinity is given by $l$ defined above, a random number in the range E-score$\pm\Delta$) was more than $1/N$, we considered the mutation non-neutral; it would be beneficial or deleterious depending on whether the binding affinity had increased or decreased, respectively.

### 4.4.6 Number of substitutions

We considered any sequence different from the ancestral sequence as a sequence that has become fixed if it ever reached a population frequency exceeding $90\%$ (a common practice in simulating populations [53, 246] to limit computational costs). Strictly speaking, fixation means an allele is present in 100% of the population. If a sequence passed the 90% threshold and dropped below this threshold more than once, we considered it as fixed only once.

## 4.5 Supplementary figures

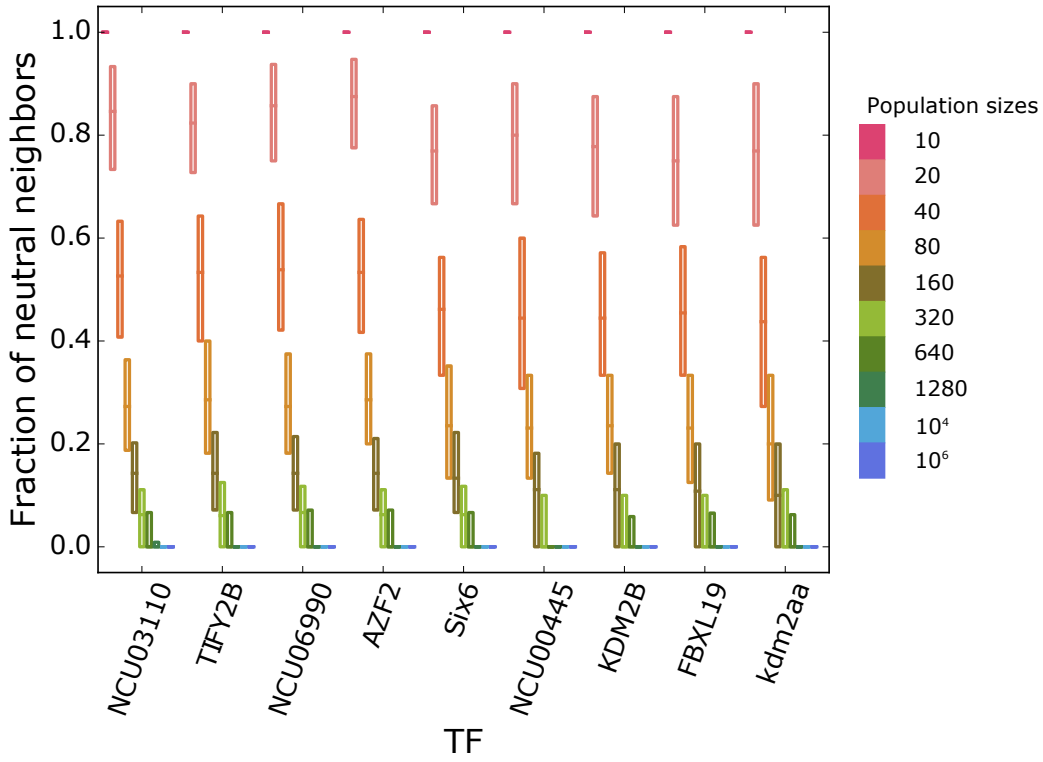


FIGURE S4.1: **Fraction of neutral single-mutation neighbors.** For each of the nine landscapes we selected all sequences in the landscape and determined the fraction of neighbors with a binding affinity difference smaller than $1/N$ for a range of population sizes (legend). In these boxplots, each box encloses the second and third quartile of the fraction of neutral neighbors among all sequences. The center line corresponds to the median. As expected, the fraction of neutral neighbors decreases with increasing population size.

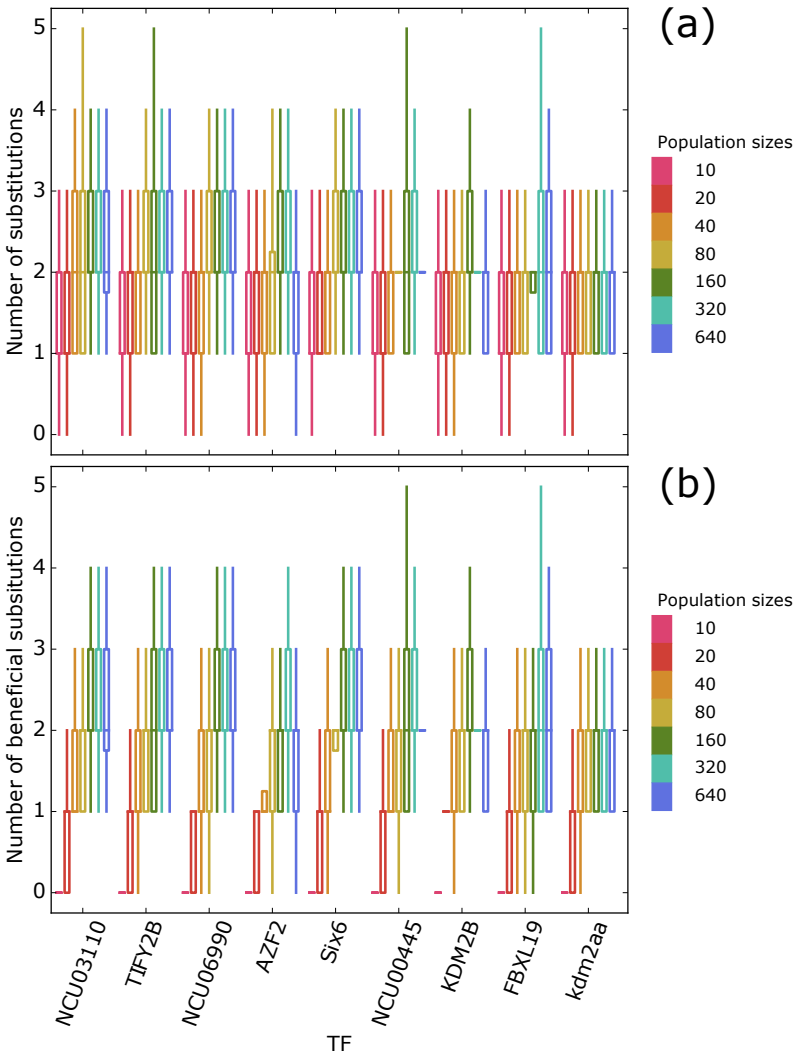


FIGURE S4.2: **Numbers of total and beneficial substitutions at $\mu = 0.001$.** The figure shows the number of **(a)** all substitutions, and **(b)** beneficial substitutions in a population, for different population sizes (color legend) and different landscapes (horizontal axis). We defined a substitution as beneficial if the sequence had a fitness increase of more than $1/N$ compared to the sequence without the mutation. Each box encloses the second and third quartiles of data from 100 replicates, the center line corresponds to the median, and the whiskers depict the minimum and maximum values obtained from any replicate, excluding outliers. Population evolution was simulated in the same way as explained in the caption of Figure 4.2.

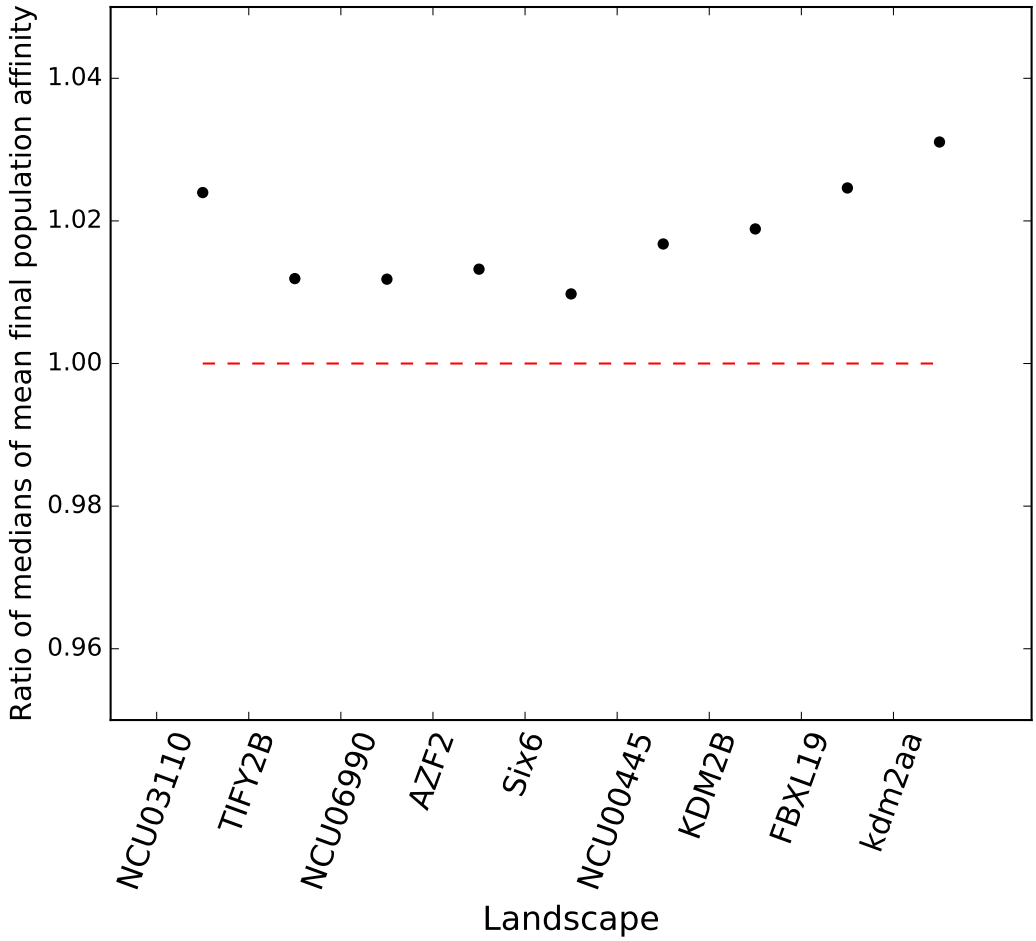


FIGURE S4.3: **How mean population affinity changes between $N = 640$ and $N = 160$, at two different mutation rates ($\mu = 0.001$ and $\mu = 0.01$).** We first divided the median of mean final population affinities for all 100 simulation replicates of $N = 640$ to that of $N = 160$ when $\mu = 0.001$. We calculated the same ratios for populations evolved at $\mu = 0.01$. We finally divided the ratios at $\mu = 0.001$ to those at $\mu = 0.01$ and plotted them as circles in this figure. The circles above 1 indicate that the difference between mean final affinity of population at $N = 640$ to $N = 160$ is larger at the smaller mutation rate of $\mu = 0.001$.

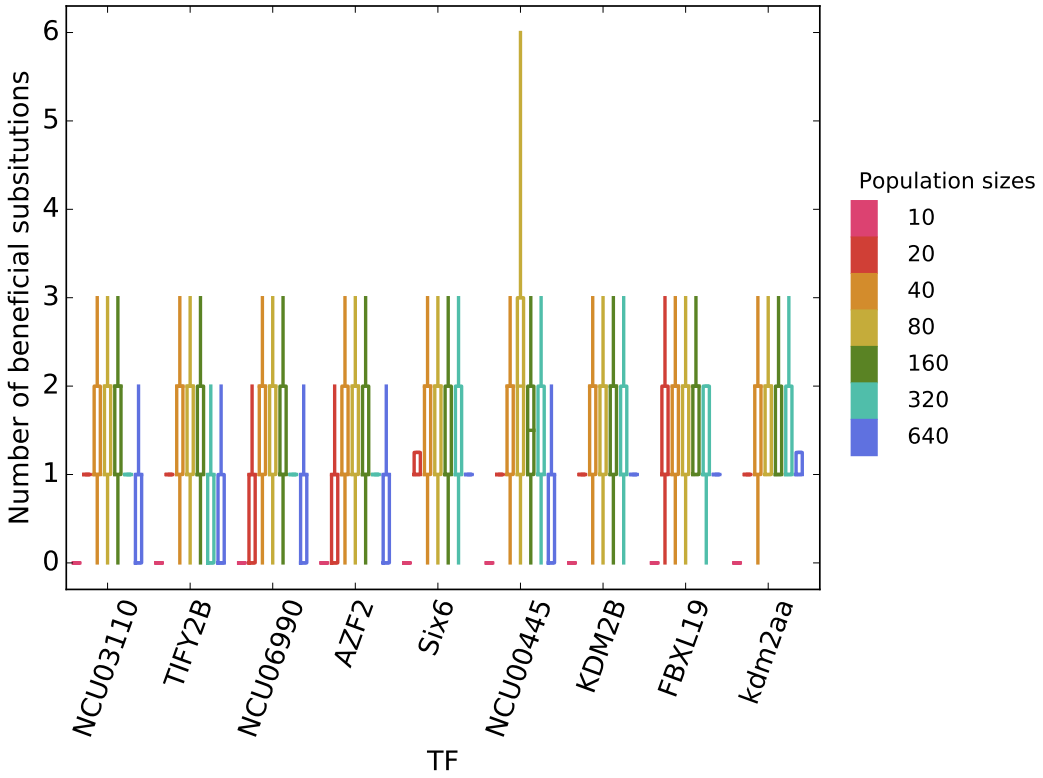


FIGURE S4.4: **Number of beneficial substitutions in all simulations at constant $\mu = 0.01$.** Each boxplot summarizes the number of beneficial substitutions in 100 simulation replicates for different landscapes (horizontal axis) and population sizes (color legend). Each box encloses the second and third quartiles of data from 100 replicates, the center line corresponds to the median, and the whiskers depict the minimum and maximum values obtained from any replicate, excluding outliers. Population evolution was simulated in the same way as explained in the caption of Figure 4.2, except that $\mu = 0.01$.

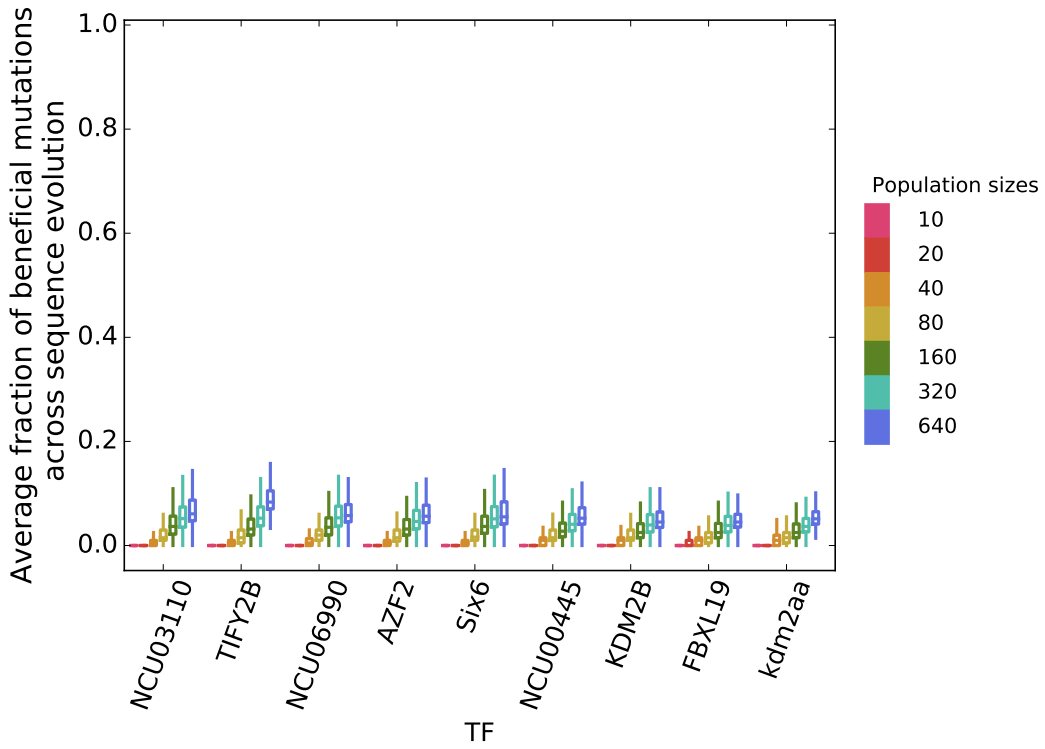


FIGURE S4.5: **Fraction of beneficial mutations at constant $\mu = 0.01$.** Each boxplot summarizes the fraction of beneficial mutations in 100 simulation replicates for different landscapes (horizontal axis) and population sizes (color legend). Each box encloses the second and third quartiles of data from 100 replicates, the center line corresponds to the median, and the whiskers depict the minimum and maximum values obtained from any replicate, excluding outliers. Population evolution was simulated in the same way as explained in the caption of Figure 4.2, except that $\mu = 0.01$.

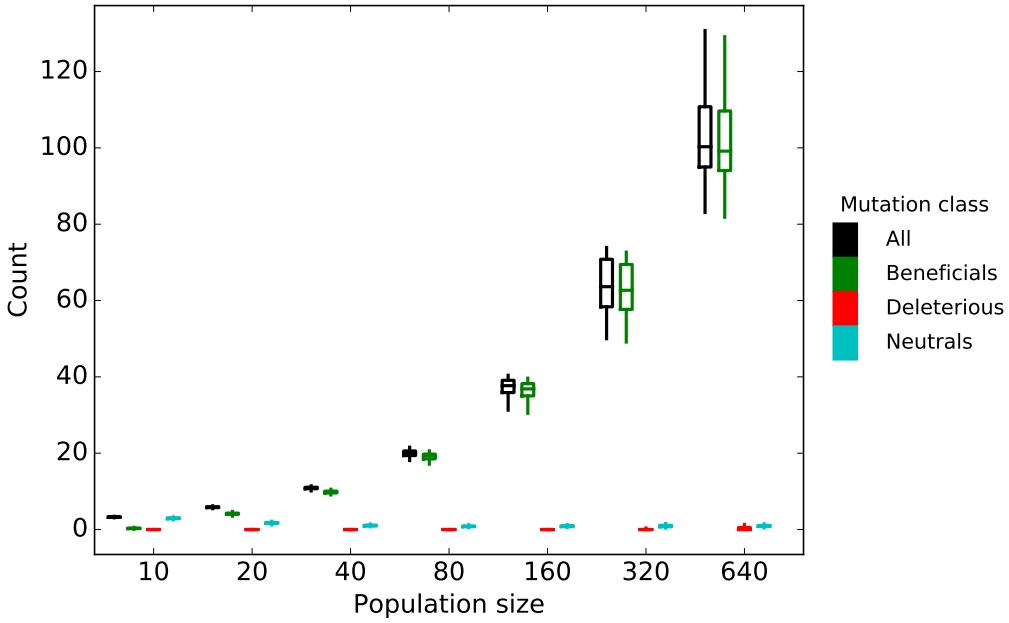


FIGURE S4.6: **More beneficial mutations coexist in larger populations evolving on the AZF2 landscape at constant $\mu = 0.1$.** Boxplots summarize mean numbers of unique total, beneficial, deleterious, and neutral mutations that coexist per generation (color legend) for populations of different sizes (horizontal axis) evolved on the AZF2 landscape. When more than one beneficial mutation is present at the same time in a population, those mutations compete for fixation (clonal interference), resulting in longer fixation time for the mutation that finally fixes in the population. We determined the effect of each mutation compared to the ancestral sequence starting the population simulation. Effects smaller than $1/N$ are neutral. Each box encloses the second and third quartiles of data from 100 replicates, the center line corresponds to the median, and the whiskers depict the minimum and maximum values obtained from any replicate, excluding outliers. Population evolution was simulated in the same way as explained in the caption of Figure 4.2, except that $\mu = 0.1$.

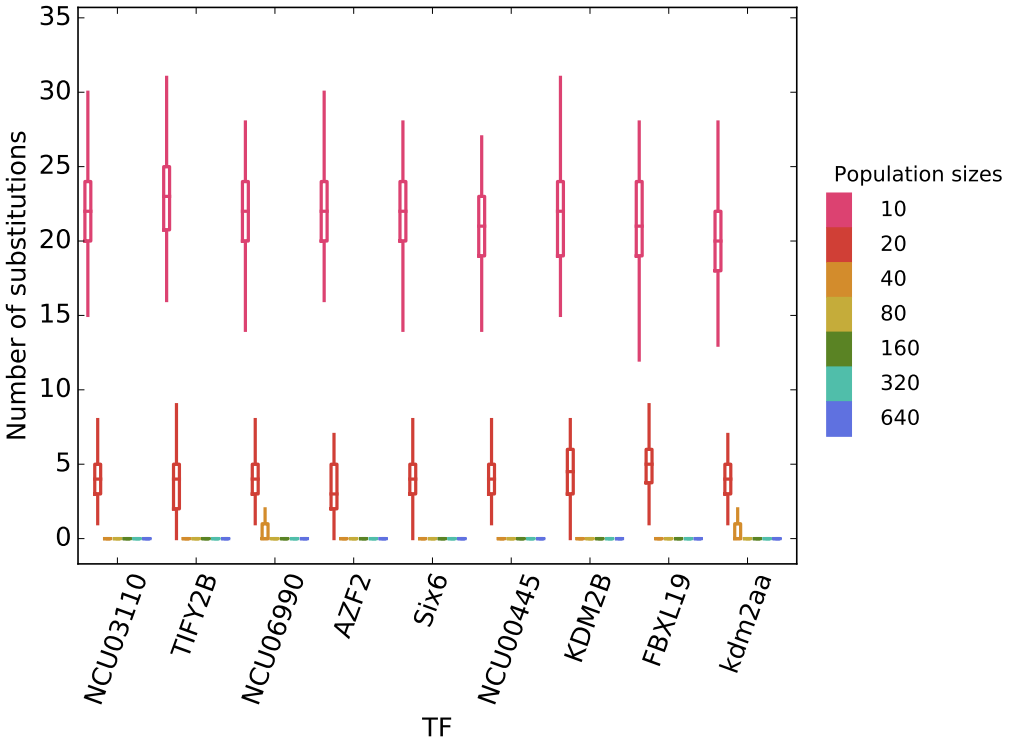


FIGURE S4.7: **Numbers of all substitutions at $\mu = 0.1$.** The figure shows the number of substitutions in a population for different population sizes (color legend) and different landscapes (horizontal axis). Each box encloses the second and third quartiles of data from 100 replicates, the center line corresponds to the median, and the whiskers depict the minimum and maximum values obtained from any replicate, excluding outliers. Population evolution was simulated in the same way as explained in the caption of Figure 4.2, except that $\mu = 0.1$.

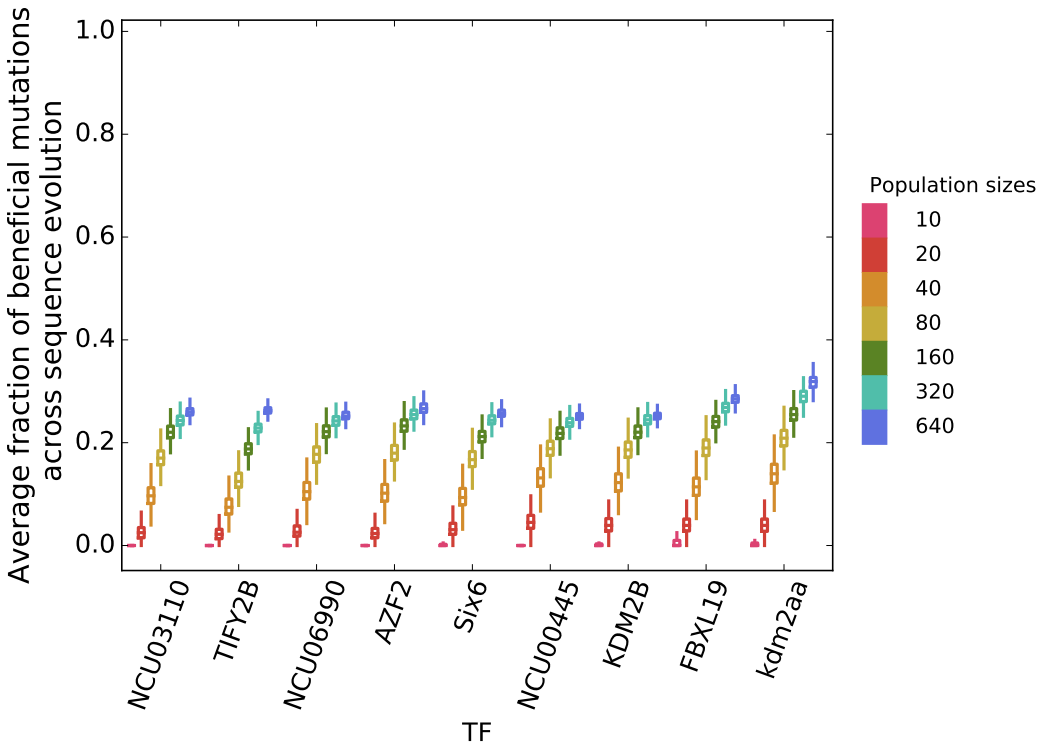


FIGURE S4.8: **Fraction of beneficial mutations at constant $\mu = 0.1$.** Each boxplot summarizes the fraction of beneficial mutations in 100 simulation replicates for different landscapes (horizontal axis) and population sizes (color legend). Each box encloses the second and third quartiles of data from 100 replicates, the center line corresponds to the median, and the whiskers depict the minimum and maximum values obtained from any replicate, excluding outliers. Population evolution was simulated in the same way as explained in the caption of Figure 4.2, except that $\mu = 0.1$.

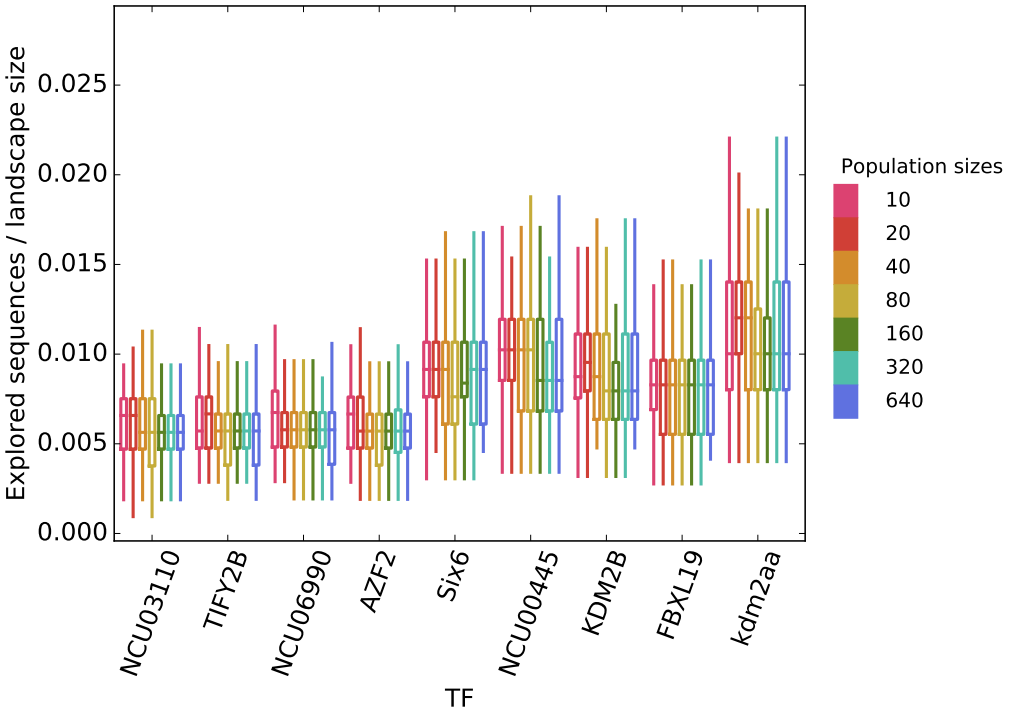


FIGURE S4.9: **Number of explored sequences across generations at constant $N\mu = 0.01$.** The figure shows the total number of unique sequences visited by a population during 1,000 generations of simulated evolution for different population sizes (color legend), normalized by the size of each landscape (horizontal axis). Each box encloses the second and third quartiles of data from 100 replicates, the center line corresponds to the median, and the whiskers depict the minimum and maximum values obtained from any replicate, excluding outliers. Population evolution was simulated in the same way as explained in the caption of Figure 4.7.

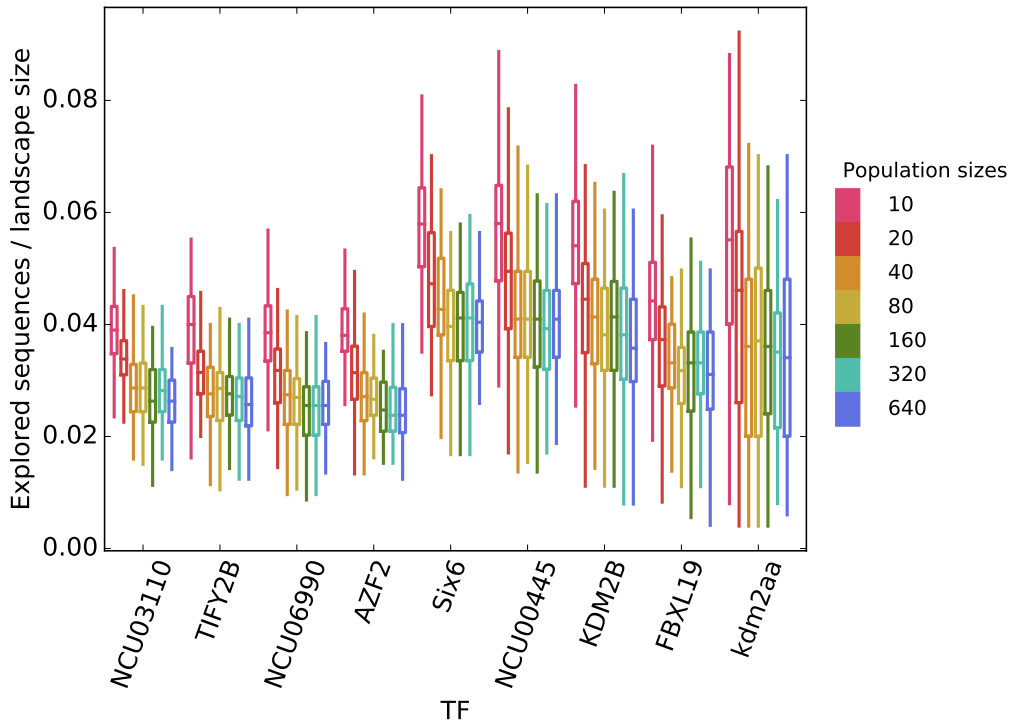


FIGURE S4.10: **Number of explored sequences across generations at constant $N\mu = 0.1$.** The figure shows the total number of unique sequences visited by a population during 1,000 generations of simulated evolution for different population sizes (color legend), normalized by the size of each landscape (horizontal axis). Each box encloses the second and third quartiles of data from 100 replicates, the center line corresponds to the median, and the whiskers depict the minimum and maximum values obtained from any replicate, excluding outliers. Population evolution was simulated in the same way as explained in the caption of Figure 4.7, except that $N\mu = 0.1$.

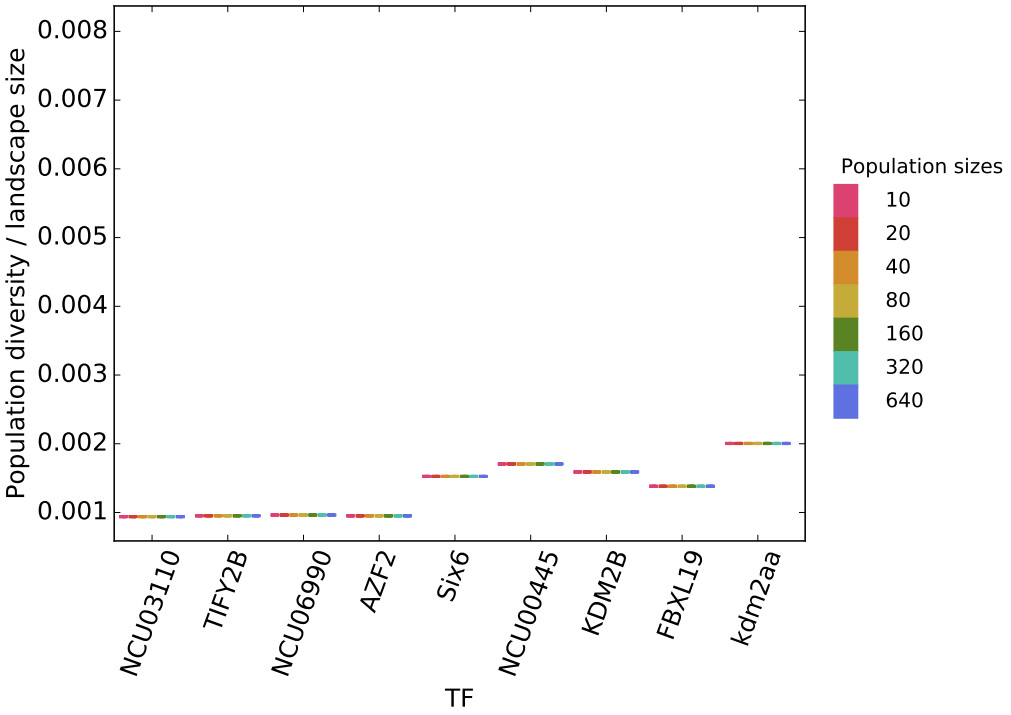


FIGURE S4.11: **Population diversity at the end of simulations at constant $N\mu = 0.01$.** The figure shows the number of unique sequences at generation 1,000 of simulated evolution for different population sizes (color legend), normalized by the size of each landscape (horizontal axis). Each box encloses the second and third quartiles of data from 100 replicates, which are smaller than the line width in this plot The center line corresponds to the median, and the whiskers depict the minimum and maximum values obtained from any replicate, excluding outliers. Population evolution was simulated in the same way as explained in the caption of Figure 4.7.

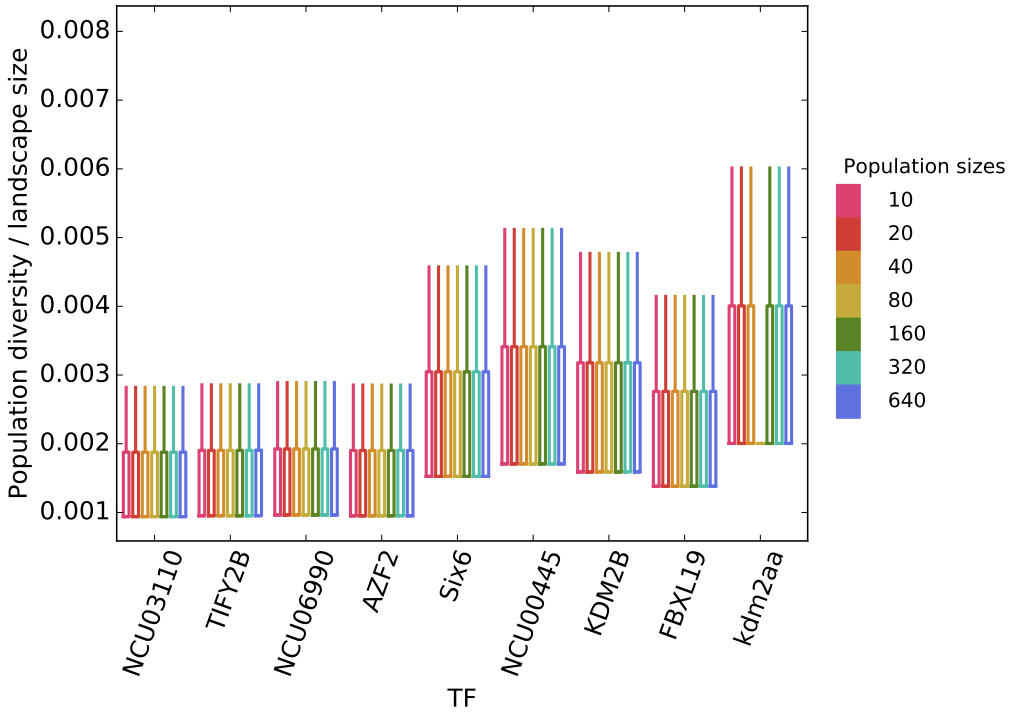


FIGURE S4.12: **Population diversity at the end of simulations at constant $N\mu = 0.1$.** The figure shows the number of unique sequences at generation 1,000 of simulated evolution for different population sizes (color legend), normalized by the size of each landscape (horizontal axis). Each box encloses the second and third quartiles of data from 100 replicates. The center line corresponds to the median, and the whiskers depict the minimum and maximum values obtained from any replicate, excluding outliers. Population evolution was simulated in the same way as explained in the caption of Figure 4.7, except that $N\mu = 0.1$.

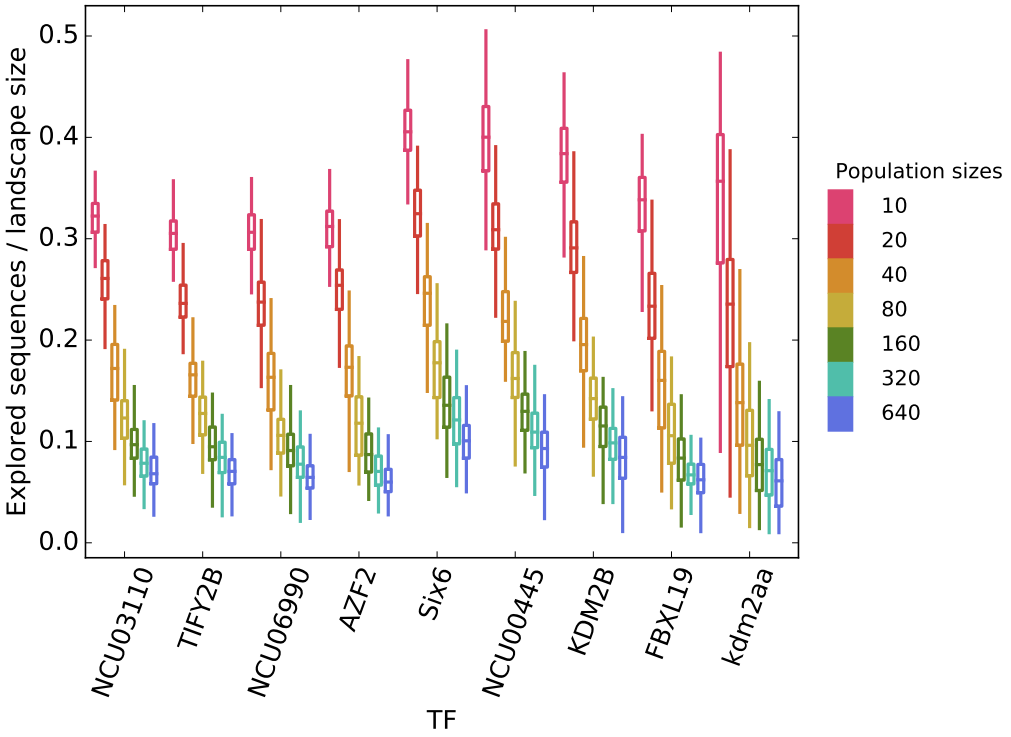


FIGURE S4.13: **Number of explored sequences across generations at constant $N\mu = 1$.** The figure shows the total number of unique sequences visited by a population during 1,000 generations of simulated evolution for different population sizes (color legend), normalized by the size of each landscape (horizontal axis). Each box encloses the second and third quartiles of data from 100 replicates, the center line corresponds to the median, and the whiskers depict the minimum and maximum values obtained from any replicate, excluding outliers. Population evolution was simulated in the same way as explained in the caption of Figure 4.7, except that $N\mu = 1$.

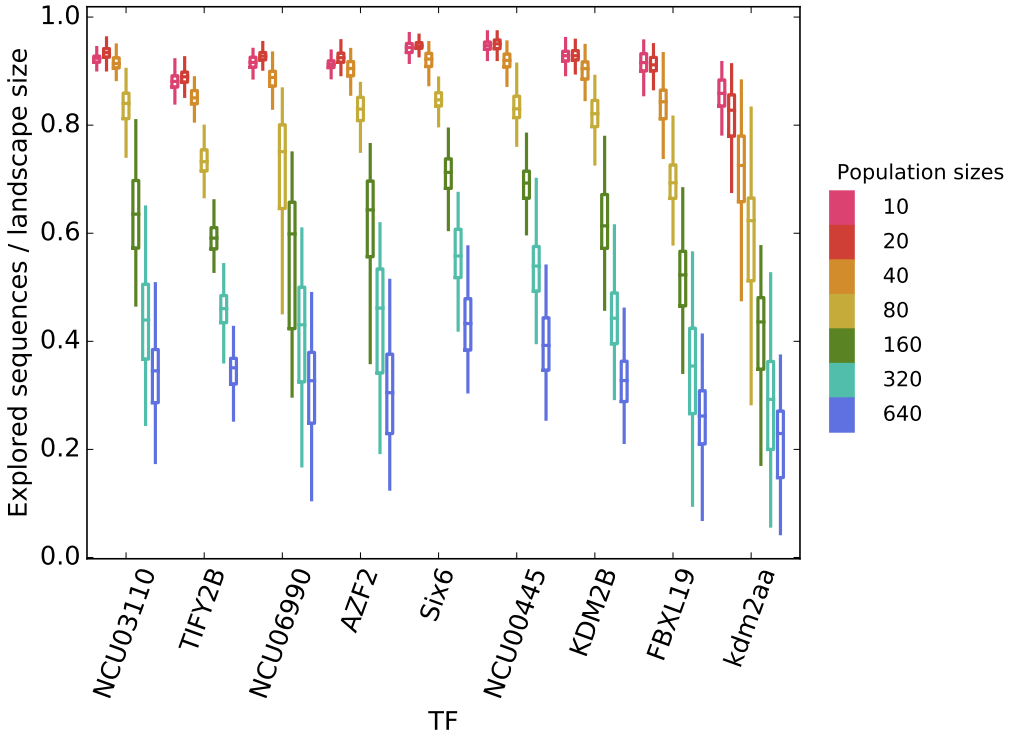


FIGURE S4.14: **Number of explored sequences across generations at constant $N\mu = 10$.** The figure shows the total number of unique sequences visited by a population during 1,000 generations of simulated evolution for different population sizes (color legend), normalized by the size of each landscape (horizontal axis). Each box encloses the second and third quartiles of data from 100 replicates, the center line corresponds to the median, and the whiskers depict the minimum and maximum values obtained from any replicate, excluding outliers. Population evolution was simulated in the same way as explained in the caption of Figure 4.7, except that $N\mu = 10$.

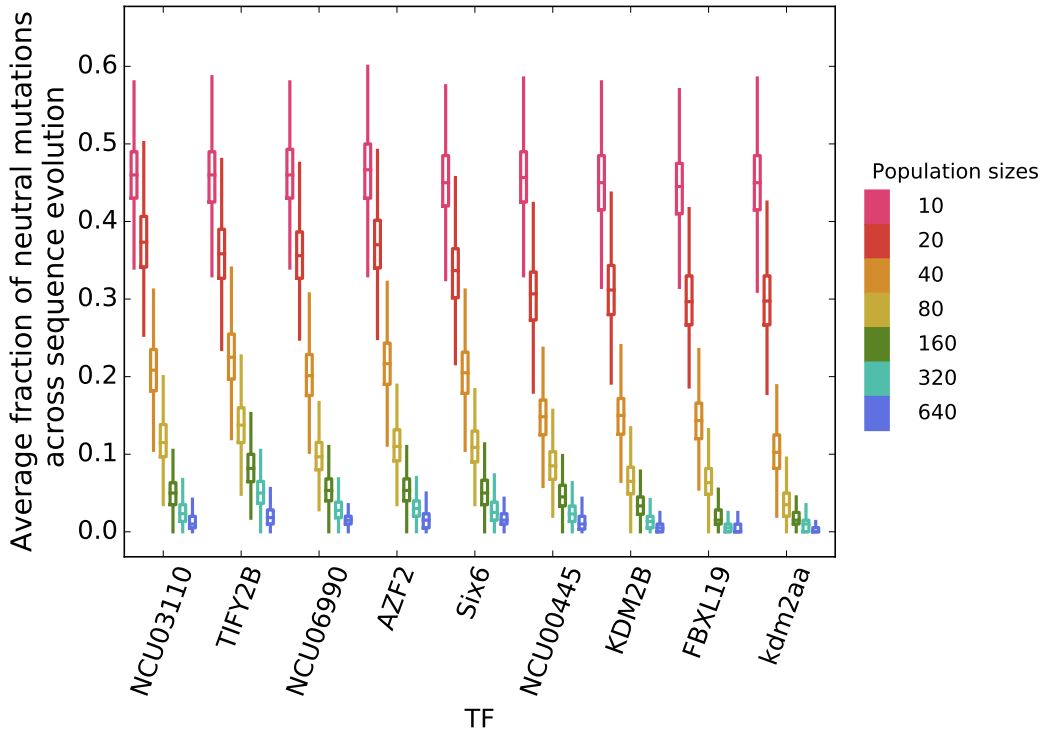


FIGURE S4.15: **Fraction of neutral mutations at constant $N\mu = 1$.** Each boxplot summarizes the fraction of neutral mutations in 100 simulation replicates for different landscapes (horizontal axis) and population sizes (color legend). Each box encloses the second and third quartiles of data from 100 replicates, the center line corresponds to the median, and the whiskers depict the minimum and maximum values obtained from any replicate, excluding outliers. Population evolution was simulated in the same way as explained in the caption of Figure 4.7, except that $N\mu = 1$.

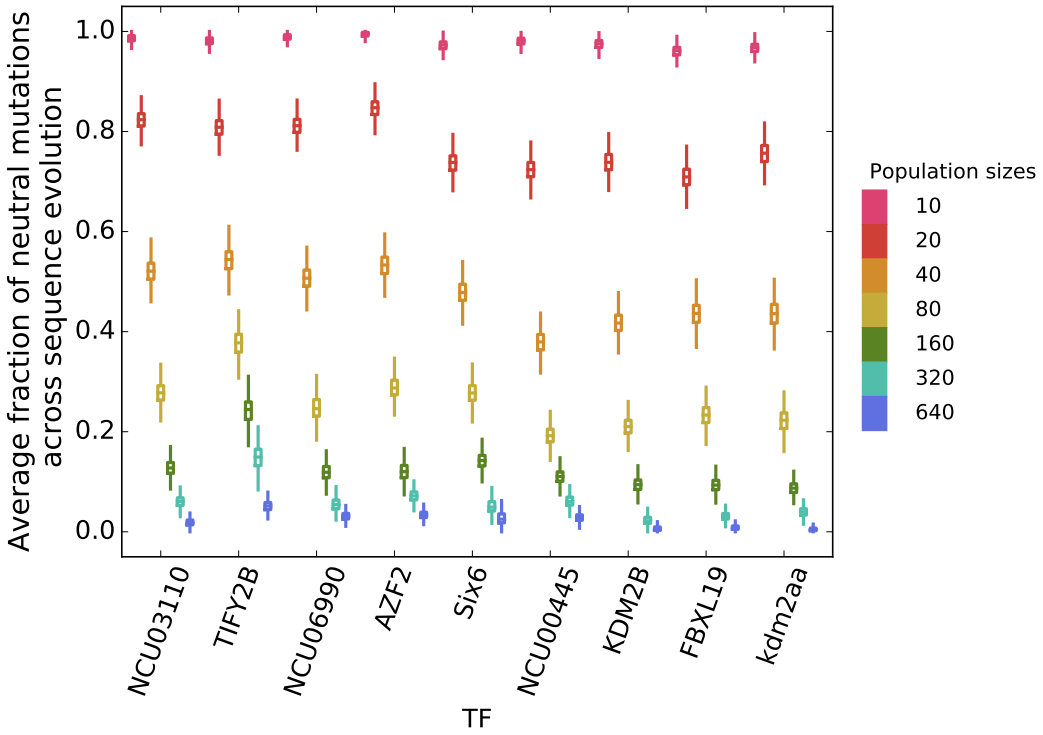


FIGURE S4.16: **Fraction of neutral mutations at constant $N\mu = 10$.** Each boxplot summarizes the fraction of neutral mutations in 100 simulation replicates for different landscapes (horizontal axis) and population sizes (color legend). Each box encloses the second and third quartiles of data from 100 replicates, the center line corresponds to the median, and the whiskers depict the minimum and maximum values obtained from any replicate, excluding outliers. Population evolution was simulated in the same way as explained in the caption of Figure 4.7, except that $N\mu = 10$.

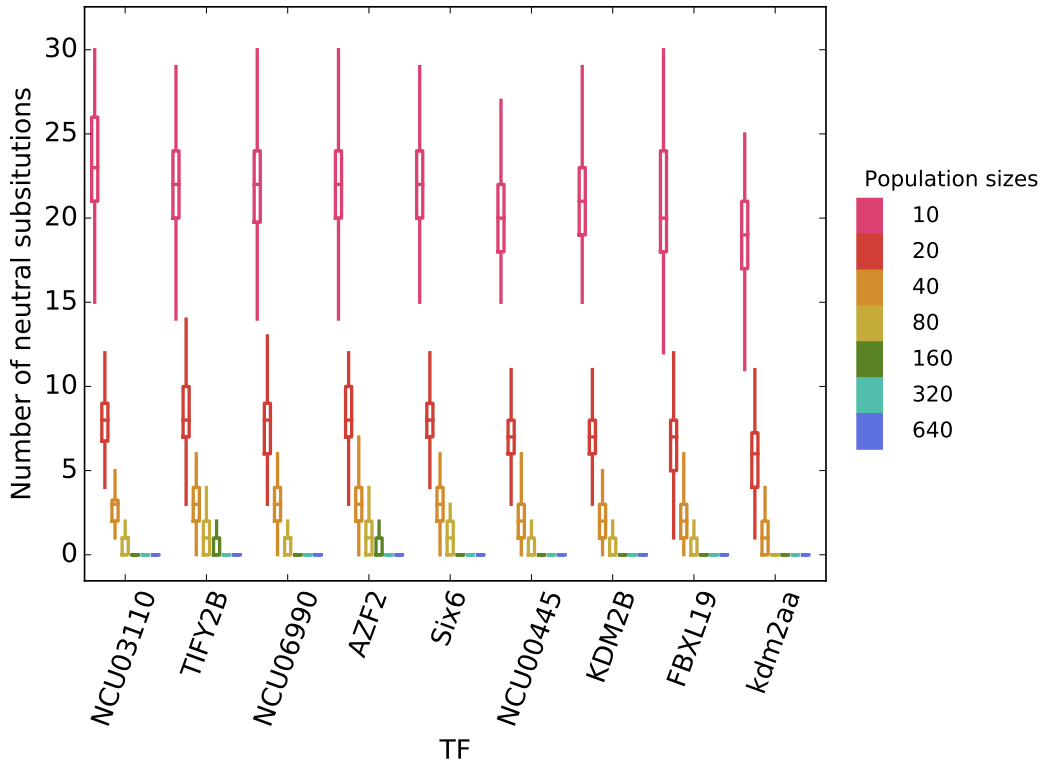


FIGURE S4.17: **Number of neutral substitutions at constant $N\mu = 1$.** Each boxplot summarizes the number of neutral substitutions in 100 simulation replicates for different landscapes (horizontal axis) and population sizes (color legend). Each box encloses the second and third quartiles of data from 100 replicates, the center line corresponds to the median, and the whiskers depict the minimum and maximum values obtained from any replicate, excluding outliers. Population evolution was simulated in the same way as explained in the caption of Figure 4.7, except that $N\mu = 1$.

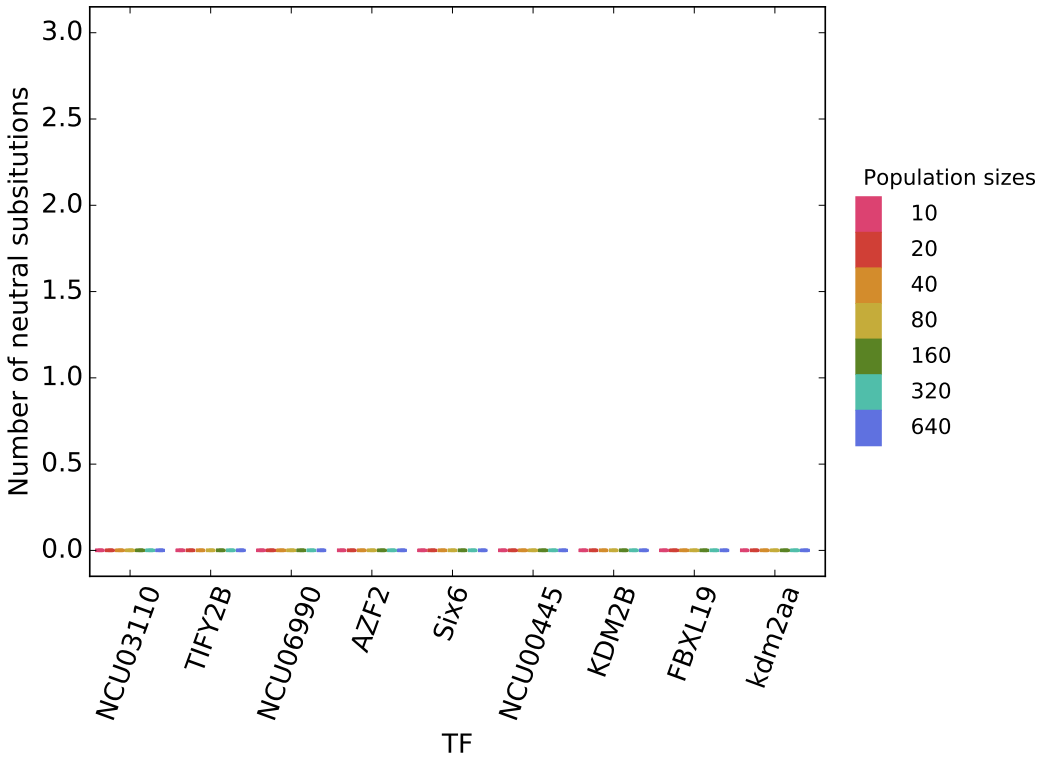


FIGURE S4.18: **Number of neutral substitutions at constant $N\mu = 10$.** Each boxplot summarizes the number of neutral substitutions in 100 simulation replicates for different landscapes (horizontal axis) and population sizes (color legend). Each box encloses the second and third quartiles of data from 100 replicates, the center line corresponds to the median, and the whiskers depict the minimum and maximum values obtained from any replicate, excluding outliers. Population evolution was simulated in the same way as explained in the caption of Figure 4.7, except that $N\mu = 10$.

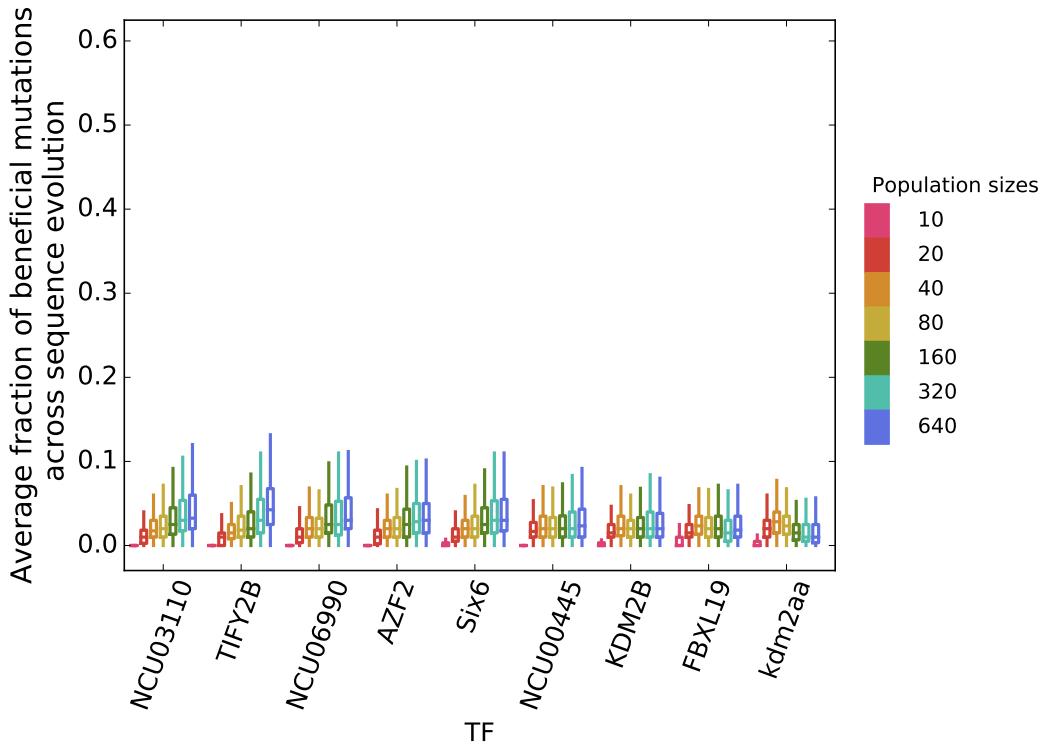


FIGURE S4.19: **Fraction of beneficial mutations at constant $N\mu = 1$.** Each boxplot summarizes the fraction of beneficial mutations in 100 simulation replicates for different landscapes (horizontal axis) and population sizes (color legend). Each box encloses the second and third quartiles of data from 100 replicates, the center line corresponds to the median, and the whiskers depict the minimum and maximum values obtained from any replicate, excluding outliers. Population evolution was simulated in the same way as explained in the caption of Figure 4.7, except that $N\mu = 1$.

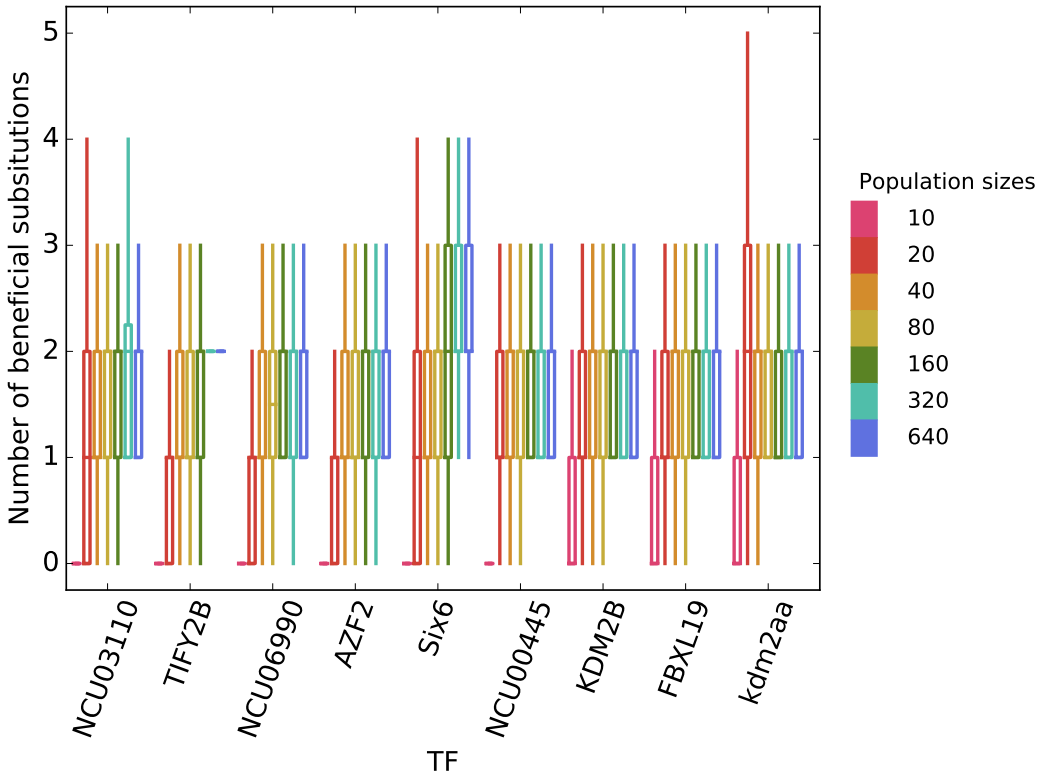


FIGURE S4.20: **Fraction of beneficial substitutions at constant $N\mu = 1$.** Each boxplot summarizes the fraction of beneficial substitutions in 100 simulation replicates for different landscapes (horizontal axis) and population sizes (color legend). Each box encloses the second and third quartiles of data from 100 replicates, the center line corresponds to the median, and the whiskers depict the minimum and maximum values obtained from any replicate, excluding outliers. Population evolution was simulated in the same way as explained in the caption of Figure 4.7, except that $N\mu = 1$.

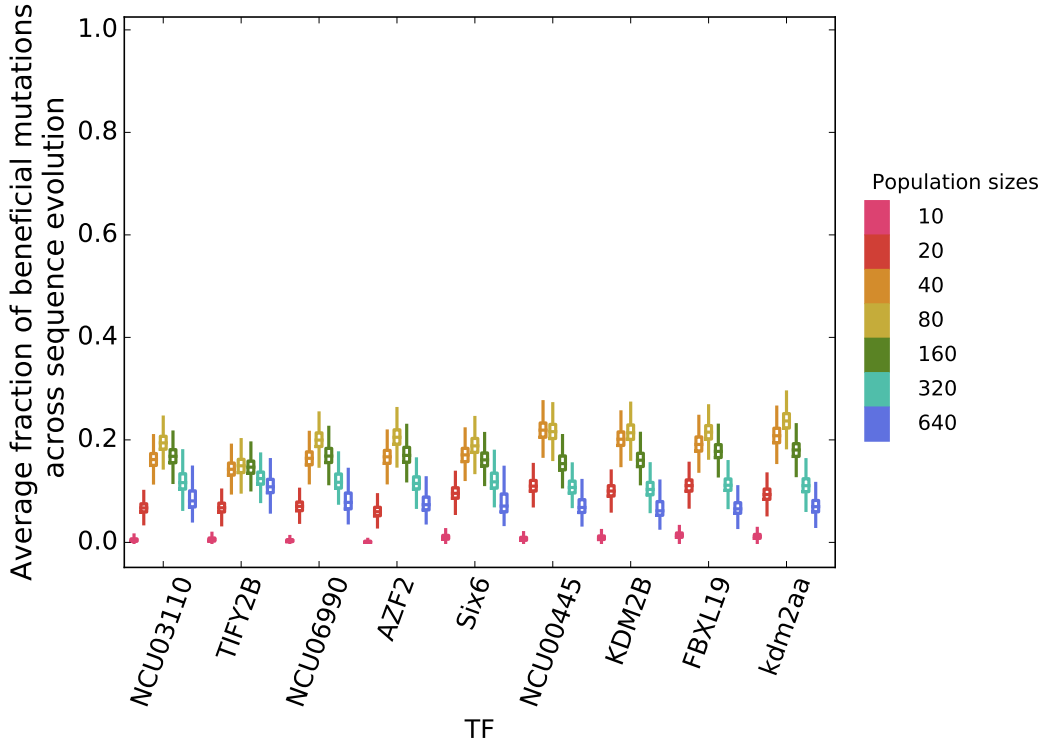


FIGURE S4.21: **Fraction of beneficial mutations at constant $N\mu = 10$.** Each boxplot summarizes the fraction of beneficial mutations in 100 simulation replicates for different landscapes (horizontal axis) and population sizes (color legend). Each box encloses the second and third quartiles of data from 100 replicates, the center line corresponds to the median, and the whiskers depict the minimum and maximum values obtained from any replicate, excluding outliers. Population evolution was simulated in the same way as explained in the caption of Figure 4.7, except that $N\mu = 10$. The likely reason why peaks occur at intermediate population sizes is that smaller populations experience fewer beneficial mutations, because selection is less efficient for them, and larger populations, having reached higher levels in the landscape, have a different distribution of fitness effects with fewer beneficial mutations. Therefore, populations at intermediate sizes experience the most beneficial mutations.

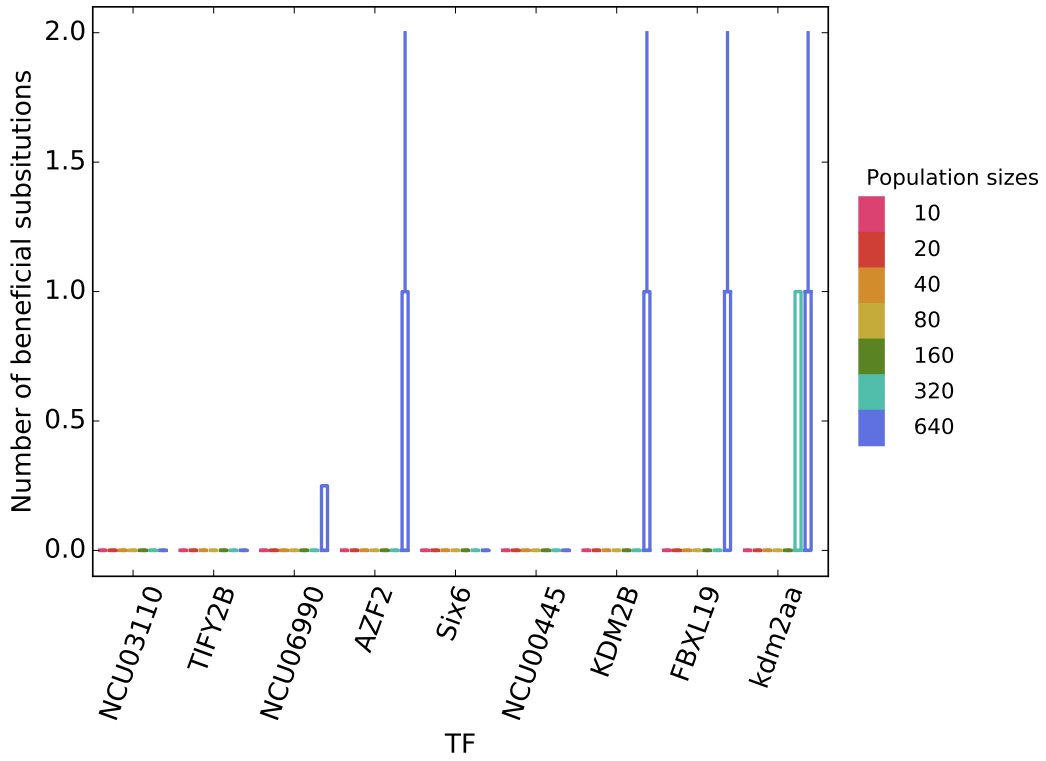


FIGURE S4.22: **Fraction of beneficial substitutions at constant $N\mu = 10$.** Each boxplot summarizes the fraction of beneficial substitutions in 100 simulation replicates for different landscapes (horizontal axis) and population sizes (color legend). Each box encloses the second and third quartiles of data from 100 replicates, the center line corresponds to the median, and the whiskers depict the minimum and maximum values obtained from any replicate, excluding outliers. Population evolution was simulated in the same way as explained in the caption of Figure 4.7, except that $N\mu = 10$.

## 4.6 Supplementary tables

TABLE S4.1: Correlations between number of peaks in each of 957 landscapes with 100 or more sequences [2] and mean affinity of populations at generation 1,000. Each row of the table represents the correlation between the mean final affinity of 100 simulation replicates for each of the landscapes at a given mutation rate, and the number of peaks in those landscapes. We normalized the mean final affinity of each population by the maximum binding affinity in the lanscape.

| **Population size** | **Mutation rate** | **Spearman's $\rho$** | **p-value** |
|---|---|---|---|
| 10 | 0.001 | -0.331 | 0 |
| 20 | 0.001 | -0.356 | 0 |
| 40 | 0.001 | -0.249 | 0 |
| 80 | 0.001 | -0.155 | 0 |
| 160 | 0.001 | -0.285 | 0 |
| 320 | 0.001 | -0.205 | 0 |
| 640 | 0.001 | -0.382 | 0 |
| 10 | 0.01 | -0.379 | 0 |
| 20 | 0.01 | -0.389 | 0 |
| 40 | 0.01 | -0.324 | 0 |
| 80 | 0.01 | -0.240 | 0 |
| 160 | 0.01 | -0.353 | 0 |
| 320 | 0.01 | -0.284 | 0 |
| 640 | 0.01 | -0.394 | 0 |
| 10 | 0.1 | -0.407 | 0 |
| 20 | 0.1 | -0.438 | 0 |
| 40 | 0.1 | -0.330 | 0 |
| 80 | 0.1 | -0.240 | 0 |
| 160 | 0.1 | -0.365 | 0 |
| 320 | 0.1 | -0.283 | 0 |
| 640 | 0.1 | -0.459 | 0 |
| 10 | 1 | -0.433 | 0 |
| 20 | 1 | -0.450 | 0 |
| 40 | 1 | -0.365 | 0 |
| 80 | 1 | -0.257 | 0 |
| 160 | 1 | -0.407 | 0 |
| 320 | 1 | -0.309 | 0 |
| 640 | 1 | -0.460 | 0 |

TABLE S4.2: Correlations between size of the global peak in each of 957 landscapes 100 or more sequences [2] and mean affinity of populations at generation 1,000. Each row of the table represents the correlation between the mean final affinity of 100 simulation replicates for each of the landscapes at a given mutation rate, and the size of the global peak in those landscapes. We normalized the mean final affinity of each population by the maximum binding affinity in the lanscape.

| **Population size** | **Mutation rate** | **Spearman's $\rho$** | **p-value** |
|---|---|---|---|
| 10 | 0.001 | 0.356 | 0 |
| 20 | 0.001 | 0.395 | 0 |
| 40 | 0.001 | 0.272 | 0 |
| 80 | 0.001 | 0.166 | 0 |
| 160 | 0.001 | 0.311 | 0 |
| 320 | 0.001 | 0.225 | 0 |
| 640 | 0.001 | 0.434 | 0 |
| 10 | 0.01 | 0.453 | 0 |
| 20 | 0.01 | 0.491 | 0 |
| 40 | 0.01 | 0.345 | 0 |
| 80 | 0.01 | 0.265 | 0 |
| 160 | 0.01 | 0.400 | 0 |
| 320 | 0.01 | 0.295 | 0 |
| 640 | 0.01 | 0.514 | 0 |
| 10 | 0.1 | 0.537 | 0 |
| 20 | 0.1 | 0.583 | 0 |
| 40 | 0.1 | 0.410 | 0 |
| 80 | 0.1 | 0.292 | 0 |
| 160 | 0.1 | 0.470 | 0 |
| 320 | 0.1 | 0.345 | 0 |
| 640 | 0.1 | 0.619 | 0 |
| 10 | 1 | 0.568 | 0 |
| 20 | 1 | 0.594 | 0 |
| 40 | 1 | 0.472 | 0 |
| 80 | 1 | 0.326 | 0 |
| 160 | 1 | 0.531 | 0 |
| 320 | 1 | 0.402 | 0 |
| 640 | 1 | 0.606 | 0 |

TABLE S4.3: Biological functions of transcription factors whose landscapes we have used in this study. We used UniProt [245] for finding the functions of the factors, and CIS-BP database [261] for finding the DNA binding domain classes.

| ID | Full name | Function | DNA binding domain class |
|---|---|---|---|
| NCU03110 | Uncharacterized | | Zinc cluster |
| TIFY2B | GATA transcription factor 24 | A transcriptional activator that binds to gene promoters. | GATA |
| NCU06990 | Uncharacterized | | Zinc cluster |
| AZF2 | Zinc finger protein AZF2 | A transcriptional repressor that prevents plant growth. | C2H2 ZF |
| Six6 | Homeobox protein SIX6 | Probably affects eye development. | |
| NCU00445 | Uncharacterized | | Zinc cluster |
| KDM2B | Lysine-specific demethylase 2B | Prevents cell growth and proliferation by repressing the transcription of ribosomal RNA genes. | CxxC |
| FBXL19 | F-box/LRR-repeat protein 19 | Part of the SCF (SKP1-CUL1-F-box protein)-type E3 ubiquitin ligase complex. | CxxC |
| kdm2aa | Uncharacterized | | CxxC |

TABLE S4.4: Correlation between the mean final binding affinity of all simulated populations, normalized by the maximum affinity in each landscape, and the number of unique sequences at generation 1,000 at constant $\mu = 0.01$. P-values are corrected for multiple testing using FDR.

| **TF** | **Spearman's $\rho$** | **p-value** |
|---|---|---|
| NCU03110 | 0.42 | 9.56E-31 |
| TIFY2B | 0.45 | 1.14E-35 |
| NCU06990 | 0.39 | 4.25E-27 |
| AZF2 | 0.37 | 1.80E-24 |
| Six6 | 0.49 | 1.38E-42 |
| NCU00445 | 0.27 | 4.97E-13 |
| KDM2B | 0.46 | 1.22E-37 |
| FBXL19 | 0.33 | 1.77E-19 |
| kdm2aa | 0.35 | 8.27E-22 |

TABLE S4.5: Correlation between the mean final binding affinity of all simulated populations, normalized by the maximum affinity in each landscape, and the number of explored sequences at constant $\mu = 0.01$. P-values are corrected for multiple testing using FDR.

| **TF** | **Spearman's $\rho$** | **p-value** |
|---|---|---|
| NCU03110 | 0.51 | 1.71E-46 |
| TIFY2B | 0.59 | 5.33E-67 |
| NCU06990 | 0.43 | 3.27E-32 |
| AZF2 | 0.43 | 1.33E-32 |
| Six6 | 0.65 | 7.76E-83 |
| NCU00445 | 0.44 | 6.11E-35 |
| KDM2B | 0.62 | 3.13E-76 |
| FBXL19 | 0.48 | 4.58E-42 |
| kdm2aa | 0.61 | 5.04E-71 |

TABLE S4.6: Correlation between the mean final binding affinity of all simulated populations, normalized by the maximum affinity in each landscape, and the number of beneficial mutations at constant $\mu = 0.01$. P-values are corrected for multiple testing using FDR.

| TF | Spearman's $\rho$ | p-value |
|---|---|---|
| NCU03110 | 0.56 | 3.75E-57 |
| TIFY2B | 0.53 | 1.09E-50 |
| NCU06990 | 0.50 | 8.72E-45 |
| AZF2 | 0.53 | 2.39E-51 |
| Six6 | 0.58 | 1.52E-64 |
| NCU00445 | 0.44 | 7.23E-34 |
| KDM2B | 0.55 | 1.99E-56 |
| FBXL19 | 0.48 | 3.92E-42 |
| kdm2aa | 0.44 | 4.48E-34 |

TABLE S4.7: Correlation between the mean final binding affinity of all simulated populations, normalized by the maximum affinity in each landscape, and the number of beneficial substitutions at constant $\mu = 0.1$. p-values are corrected for multiple testing using FDR.

| TF | Spearman's $\rho$ | p-value |
|---|---|---|
| NCU03110 | 0.43 | 3.17E-32 |
| TIFY2B | 0.53 | 1.86E-51 |
| NCU06990 | 0.62 | 2.12E-74 |
| AZF2 | 0.45 | 2.16E-36 |
| Six6 | 0.39 | 5.06E-26 |
| NCU00445 | 0.36 | 8.74E-23 |
| KDM2B | 0.36 | 4.78E-23 |
| FBXL19 | 0.32 | 7.02E-18 |
| kdm2aa | 0.54 | 4.86E-54 |

TABLE S4.8: Delta ($\Delta$) values used in our simulations as a measure of noise in measured affinity E-scores, taken from [2].

| TF name | Delta |
|---|---|
| NCU03110 | 0.024419 |
| TIFY2B | 0.024981 |
| NCU06990 | 0.028733 |
| AZF2 | 0.022419 |
| Six6 | 0.024746 |
| NCU00445 | 0.031016 |
| KDM2B | 0.028908 |
| FBXL19 | 0.028274 |
| kdm2aa | 0.028421 |

# Bibliography


[1] Rebecca R Ackermann and James M Cheverud. "Detecting genetic drift versus selection in human evolution". In: *Proceedings of the National Academy of Sciences of the United States of America* 101.52 (2004), pp. 17946–17951. ISSN: 0027-8424. DOI: 10.1073/pnas.0405919102.

[2] José Aguilar-Rodríguez, Joshua L Payne, and Andreas Wagner. "A thousand empirical adaptive landscapes and their navigability". In: *Nature Ecology & Evolution* 1.2 (2017), p. 0045. ISSN: 2397-334X. DOI: 10.1038/s41559-016-0045.

[3] Hiroshi Akashi, Naoki Osada, and Tomoko Ohta. "Weak selection and protein evolution". In: *Genetics* 192.1 (2012), pp. 15–31. ISSN: 0016-6731. DOI: 10.1534/genetics.112.140178.

[4] Alexander Altland et al. "Rare events in population genetics: stochastic tunneling in a two-locus model with recombination". In: *Physical Review Letters* 106.8 (2011), p. 088101. ISSN: 1079-7114. DOI: 10.1103/PhysRevLett.106.088101.

[5] Lauren W Ancel and Walter Fontana. "Plasticity, evolvability, and modularity in RNA". In: *The Journal of Experimental zoology* 288.3 (2000), pp. 242–83. ISSN: 0022-104X.

[6] S Anderson et al. "Sequence and organization of the human mitochondrial genome". In: *Nature* 290.5806 (1981), pp. 457–65. ISSN: 0028-0836.

[7] Mirela Andronescu et al. "Efficient parameter estimation for RNA secondary structure prediction". In: *Bioinformatics* 23.13 (2007), pp. i19–28. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btm223.

[8] M Madan Babu et al. "Structure and evolution of transcriptional regulatory networks". In: *Current Opinion in Structural Biology* 14.3 (2004), pp. 283–91. ISSN: 0959-440X. DOI: 10.1016/j.sbi.2004.05.004.

[9] Doris Bachtrog. "Similar rates of protein adaptation in Drosophila miranda and *D. melanogaster*, two species with different current effective population sizes". In: *BMC Evolutionary Biology* 8.1 (2008), p. 334. ISSN: 1471-2148. DOI: 10.1186/1471-2148-8-334.

[10] Gwenael Badis et al. "Diversity and complexity in DNA recognition by transcription factors". In: *Science (New York, N.Y.)* 324.5935 (2009), pp. 1720–3. ISSN: 1095-9203. DOI: 10.1126/science.1162327.

[11] Gwenael Badis et al. "Diversity and complexity in DNA recognition by transcription factors". In: *Science* 324.5935 (2009), pp. 1720–3. ISSN: 1095-9203. DOI: 10.1126/science.1162327.

[12] Susan F Bailey et al. "What drives parallel evolution?" In: *BioEssays* 39.1 (2017), e201600176. ISSN: 02659247. DOI: 10.1002/bies.201600176.

[13] Albert-László Barabási and Márton Pósfai. *Network science*. 1st edition. Cambridge University Press, 2016. ISBN: 978-1107076266.

[14] Jeffrey E Barrick et al. "Genome evolution and adaptation in a long-term experiment with *Escherichia coli*". In: *Nature* 461.7268 (2009), pp. 1243–1247. ISSN: 1476-4687. DOI: 10.1038/nature08480.

[15] Jordi Bascompte, Pedro Jordano, and Jens M Olesen. "Asymmetric coevolutionary networks facilitate biodiversity maintenance". In: *Science* 312.5772 (2006), pp. 431–3. ISSN: 1095-9203. DOI: 10.1126/science.1123412.

[16] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. "Gephi: an open source software for exploring and manipulating networks". In: *International AAAI Conference on Weblogs and Social Media* (2009).

[17] Yoav Benjamini and Yosef Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 57.1 (1995), pp. 289 –300. ISSN: 00359246. DOI: 10.2307/2346101. arXiv: 95/57289 [0035-9246].

[18] Michael F Berger and Martha L. Bulyk. "Protein binding microarrays (PBMs) for rapid, high-throughput characterization of the sequence specificities of DNA binding proteins". In: *Methods in Molecular Biology* 338.617 (2006), pp. 245–60. ISSN: 1064-3745. DOI: 10.1385/1-59745-097-9:245.

[19] Michael F Berger and Martha L Bulyk. "Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors". In: *Nature Protocols* 4.3 (2009), pp. 393–411. ISSN: 1754-2189. DOI: 10.1038/nprot.2008.195.

[20] Michael F Berger et al. "Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities". In: *Nature Biotechnology* 24.11 (2006), pp. 1429–1435. ISSN: 1087-0156. DOI: 10.1038/nbt1246.

[21] C William Birky and J Bruce Walsh. "Effects of linkage on rates of molecular evolution". In: *Proceedings of the National Academy of Sciences of the United States of America* 85.17 (1988), pp. 6414–8. ISSN: 0027-8424.

[22] Zoë Birtle, Leo Goodstadt, and Chris Ponting. "Duplication and positive selection among hominin-specific PRAME genes". In: *BMC Genomics* 6 (2005), p. 120. ISSN: 1471-2164. DOI: 10.1186/1471-2164-6-120.

[23] Caitlin D Blaskewicz, Jeffrey Pudney, and Deborah J Anderson. "Structure and function of intercellular junctions in human cervical and vaginal mucosal epithelia". In: *Biology of Reproduction* 85.1 (2011), pp. 97–104. ISSN: 1529-7268. DOI: 10.1095/biolreprod.110.090423.

[24] Adrian Bondy and U S R Murty. *Graph theory (Graduate Texts in Mathematics)*. Springer, 2008. ISBN: 978-1846289699.

[25] Stephanie Boue, Ivica Letunic, and Peer Bork. "Alternative splicing and evolution". In: *BioEssays : News and Reviews in Molecular, Cellular and Developmental Biology* 25.11 (2003), pp. 1031–4. ISSN: 0265-9247. DOI: 10.1002/bies.10371.

[26] Débora Y C Brandt et al. "Mapping bias overestimates reference allele frequencies at the HLA genes in the 1000 genomes project phase I data". In: *G3:Genes | Genomes | Genetics* 5.5 (2015), pp. 931–941. ISSN: 2160-1836. DOI: 10.1534/g3.114.015784.

[27] Elena B M Breidenstein et al. "Complex ciprofloxacin resistome revealed by screening a *Pseudomonas aeruginosa* mutant library for altered susceptibility". In: *Antimicrobial agents and chemotherapy* 52.12 (2008), pp. 4486–91. ISSN: 1098-6596. DOI: 10.1128/AAC.00222-08.

[28] Paulo R A Campos and L M Wahl. "The effects of population bottlenecks on clonal interference, and the adaptation effective population size". In: *Evolution* 63.4 (2009), pp. 950–958. ISSN: 00143820. DOI: 10.1111/j.1558-5646.2008.00595.x.

[29] Paulo R A Campos and Lindi M Wahl. "The adaptation rate of asexuals: deleterious mutations, clonal interference and population bottlenecks". In: *Evolution* 64.7 (2010), pp. 1973–1983. ISSN: 00143820. DOI: 10.1111/j.1558-5646.2010.00981.x.

[30] Rebecca L Cann, Mark Stoneking, and Allan C Wilson. "Mitochondrial DNA and human evolution". In: *Nature* 325.6099 (1987), pp. 31–36. ISSN: 0028-0836. DOI: 10.1038/325031a0.

[31] Thomas M Carlile et al. "Pseudouridine profiling reveals regulated mRNA pseudouridylation in yeast and human cells". In: *Nature* 515.7525 (2014), pp. 143–146. ISSN: 0028-0836. DOI: 10.1038/nature13802.

[32] Todd A Castoe et al. "Evidence for an ancient adaptive episode of convergent molecular evolution". In: *Proceedings of the National Academy of Sciences of the United States of America* 106.22 (2009), pp. 8986–91. ISSN: 0027-8424. DOI: 10.1073/pnas.0900233106.

[33] T W Chang. "Binding of cells to matrixes of distinct antibodies coated on solid surface". In: *Journal of Immunological Methods* 65.1-2 (1983), pp. 217–23. ISSN: 0022-1759.

[34] Brian Charlesworth. "Effective population size". In: *Current Biology* 12.21 (2002), R716–R717. ISSN: 09609822. DOI: 10.1016/S0960-9822(02)01244-7.

[35] Brian Charlesworth. "Effective population size and patterns of molecular evolution and variation". In: *Nature Reviews. Genetics* 10.3 (2009), pp. 195–205. ISSN: 1471-0064. DOI: 10.1038/nrg2526.

[36] Jane Charlesworth and Adam Eyre-Walker. "The rate of adaptive evolution in enteric bacteria". In: *Molecular Biology and Evolution* 23.7 (2006), pp. 1348–56. ISSN: 0737-4038. DOI: 10.1093/molbev/msk025.

[37] Hua Chen, Nick Patterson, and David Reich. "Population differentiation as a test for selective sweeps". In: *Genome Research* 20.3 (2010), pp. 393–402. ISSN: 1549-5469. DOI: 10.1101/gr.100545.109.

[38] Reshmi Chowdhury et al. "Genetic analysis of variation in human meiotic recombination". In: *PLoS Genetics* 5.9 (2009). Ed. by Gregory P. Copenhaver, e1000648. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1000648.

[39] Pascal-Antoine Christin, Daniel M Weinreich, and Guillaume Besnard. "Causes and evolutionary significance of genetic convergence". In: *Trends in Genetics : TIG* 26.9 (2010), pp. 400–5. ISSN: 0168-9525. DOI: 10.1016/j.tig.2010.06.005.

[40] S Ciliberti, O C Martin, and A Wagner. "Innovation and robustness in complex regulatory gene networks". In: *Proceedings of the National Academy of Sciences of the United States of America* 104.34 (2007), pp. 13591–6. ISSN: 0027-8424. DOI: 10.1073/pnas.0705396104.

[41] Stefano Ciliberti, Olivier C. Martin, and Andreas Wagner. "Robustness can evolve gradually in complex regulatory gene networks with varying topology". In: *PLoS Computational Biology* 3.2 (2007), e15. ISSN: 1553-734X. DOI: 10.1371/journal.pcbi.0030015.

[42] Suzanne Clancy. "RNA functions". In: *Nature Education* 1.1 (2008), p. 102.

[43] Pamela F Colosimo et al. "Widespread parallel evolution in sticklebacks by repeated fixation of Ectodysplasin alleles". In: *Science* 307.5717 (2005), pp. 1928–33. ISSN: 1095-9203. DOI: 10.1126/science.1107239.

[44] Tim F Cooper. "Recombination speeds adaptation by reducing competition between beneficial mutations in populations of *Escherichia coli*". In: *PLoS Biology* 5.9 (2007), e225. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.0050225.

[45] Heather J. Cordell. "Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans". In: *Human Molecular Genetics* 11.20 (2002), pp. 2463–2468. ISSN: 14602083. DOI: 10.1093/hmg/11.20.2463.

[46] Matthew C Cowperthwaite, J J Bull, and Lauren Ancel Meyers. "Distributions of beneficial fitness effects in RNA". In: *Genetics* 170.4 (2005), pp. 1449–57. ISSN: 0016-6731. DOI: 10.1534/genetics.104.039248.

[47] Bernard J Crespi and Kyle Summers. "Positive selection in the evolution of cancer". In: *Biological Reviews of the Cambridge Philosophical Society* 81 (2006), pp. 407–424. ISSN: 1464-7931. DOI: 10.1017/S1464793106007056.

[48] Gabor Csardi and Tamas Nepusz. "The igraph software package for complex network research". In: *InterJournal* Complex Sy (2006), p. 1695.

[49] Anna Czerwoniec et al. "MODOMICS: a database of RNA modification pathways. 2008 update". In: *Nucleic Acids Research* 37.Database (2009), pp. D118–D121. ISSN: 0305-1048. DOI: 10.1093/nar/gkn710.

[50] Giovanni Marco Dall'Olio et al. "Human genome variation and the concept of genotype networks". In: *PLoS One* 9.6 (2014). Ed. by David Caramelli, e99424. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0099424.

[51] Petr Danecek et al. "The variant call format and VCFtools". In: *Bioinformatics* 27.15 (2011), pp. 2156–2158. DOI: 10.1093/bioinformatics/btr330.

[52] K T J Davies et al. "Parallel signatures of sequence evolution among hearing genes in echolocating mammals: an emerging model of genetic convergence". In: *Heredity* 108.5 (2012), pp. 480–489. ISSN: 0018-067X. DOI: 10.1038/hdy.2011.119.

[53] Michael M Desai and Daniel S Fisher. "Beneficial mutation-selection balance and the effect of linkage on positive selection". In: *Genetics* 176.3 (2007), pp. 1759–1798. ISSN: 00166731. DOI: 10.1534/genetics.106.067678. eprint: 0612016 (q-bio).

[54] Michael M Desai, Daniel S Fisher, and Andrew W Murray. "The speed of evolution and maintenance of variation in asexual populations". In: *Current Biology* 17.5 (2007), pp. 385–394. ISSN: 09609822. DOI: 10.1016/j.cub.2007.01.072.

[55] Alivia Dey et al. "Molecular hyperdiversity defines populations of the nematode *Caenorhabditis brenneri*". In: *Proceedings of the National Academy of Sciences* 110.27 (2013), pp. 11056–11060. ISSN: 0027-8424. DOI: 10.1073/pnas.1303057110.

[56] J B Dodgson and R D Wells. "Synthesis and thermal melting behavior of oligomer-polymer complexes containing defined lengths of mismatched dA-dG and dG-dG nucleotides". In: *Biochemistry* 16.11 (1977), pp. 2367–74. ISSN: 0006-2960.

[57] Russel F Doolittle. "Convergent evolution: The need to be explicit". In: *Trends in Biochemical Sciences* 19.January (1994), pp. 15–18. ISSN: 09680004. DOI: 10.1016/0968-0004(94)90167-8.

[58] Jennifer A Doudna. "Structural genomics of RNA". In: *Nature Structural Biology* 7 Suppl (2000), pp. 954–6. ISSN: 1072-8368. DOI: 10.1038/80729.

[59] John W. Drake. "Avoiding dangerous missense: thermophiles display especially low mutation rates". In: *PLoS Genetics* 5.6 (2009). ISSN: 15537390. DOI: 10.1371/journal.pgen.1000520.

[60] T Gregory Drummond, Michael G Hill, and Jacqueline K Barton. "Electrochemical DNA sensors". In: *Nature Biotechnology* 21.10 (2003), pp. 1192–1199. ISSN: 1087-0156. DOI: 10.1038/nbt873.

[61] Richard M. R.M. M Durbin et al. "A map of human genome variation from population-scale sequencing". In: *Nature* 467.7319 (2010), pp. 1061–1073. ISSN: 1476-4687. DOI: 10.1038/nature09534.

[62] Sean R Eddy and Richard Durbin. "RNA sequence analysis using covariance models". In: *Nucleic Acids Research* 22.11 (1994), pp. 2079–88. ISSN: 0305-1048.

[63] Roger Ekins and Frederick W Chu. "Microarrays: their origins and applications". In: *Trends in Biotechnology* 17.6 (1999), pp. 217–8. ISSN: 0167-7799.

[64] Santiago F Elena and Richard E Lenski. "Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation". In: *Nature Reviews. Genetics* 4.6 (2003), pp. 457–69. ISSN: 1471-0056. DOI: 10.1038/nrg1088.

[65] Jason R Epstein, Israel Biran, and David R Walt. "Fluorescence-based nucleic acid detection and microarrays". In: *Analytica Chimica Acta* 469.1 (2002), pp. 3–36. ISSN: 00032670. DOI: 10.1016/S0003-2670(02)00030-2.

[66] Leonhard Euler. "Solutio Problemat is ad Geometriam Situs Pertinentis". In: *Commentarii Academiae Scientiarum Imperialis Petropolitanae* 8 (1741), pp. 128–140.

[67] Ralph Evans. "Niche expansion in bacteria: can infectious gene exchange affect the rate of evolution?" In: *Genetics* 113.3 (1986), pp. 775–95. ISSN: 0016-6731.

[68] Adam Eyre-Walker and Peter D Keightley. "The distribution of fitness effects of new mutations". In: *Nature Reviews. Genetics* 8.8 (2007), pp. 610–8. ISSN: 1471-0056. DOI: 10.1038/nrg2146.

[69] Kelly M Fahrbach et al. "Differential binding of IgG and IgA to mucus of the female reproductive tract". In: *PloS One* 8.10 (2013), e76176. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0076176.

[70] Justin C Fay and Chung I Wu. "Hitchhiking under positive Darwinian selection". In: *Genetics* 155.3 (2000), pp. 1405–1413. ISSN: 0016-6731.

[71] Milan Fedurco et al. "BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies". In: *Nucleic Acids Research* 34.3 (2006), e22. ISSN: 1362-4962. DOI: 10.1093/nar/gnj023.

[72] Walter Fiers et al. "Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene". In: *Nature* 260.5551 (1976), pp. 500–7. ISSN: 0028-0836.

[73] Robert Andrew Foley and Roger Lewin. *Principles of Human Evolution*. 2nd Editio. Wiley-Blackwell, 2013.

[74] Walter Fontana and Peter Schuster. "Continuity in evolution: on the nature of transitions". In: *Science (New York, N.Y.)* 280.5368 (1998), pp. 1451–5. ISSN: 0036-8075. DOI: 10.1126/science.280.5368.1451.

[75] Richard Frankham. "Effective population size/adult population size ratios in wildlife: a review". In: *Genetical Research* 89.5-6 (2007), pp. 491–503. DOI: 10.1017/S0016672308009695.

[76] I R Franklin and R Frankham. "How large must populations be to retain evolutionary potential?" In: *Animal Conservation* 1.1 (1998), pp. 69–70. ISSN: 1367-9430. DOI: 10.1111/j.1469-1795.1998.tb00228.x.

[77] Kelly A Frazer et al. "A second generation human haplotype map of over 3.1 million SNPs". In: *Nature* 449.7164 (2007), pp. 851–861. ISSN: 1476-4687. DOI: 10.1038/nature06258.

[78] A Furka et al. "General method for rapid synthesis of multicomponent peptide mixtures". In: *International Journal of Peptide and Protein Research* 37.6 (1991), pp. 487–93. ISSN: 0367-8377.

[79] Sylvain Gandon et al. "Forecasting epidemiological and evolutionary dynamics of infectious diseases". In: *Trends in Ecology & Evolution* 31.10 (2016), pp. 776–88. ISSN: 1872-8383. DOI: 10.1016/j.tree.2016.07.010.

[80] Philippe Gayral et al. "Reference-free population genomics from next-generation transcriptome data and the vertebrate–invertebrate gap". In: *PLoS Genetics* 9.4 (2013). Ed. by John J. Welch, e1003457. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1003457.

[81] Philip J Gerrish and Richard E Lenski. "The fate of competing beneficial mutations in an asexual population". In: *Genetica* 102-103.1-6 (1998), pp. 127–44. ISSN: 0016-6707. DOI: 10.1023/A:1017067816551.

[82] John H Gillespie. "A simple stochastic gene substitution model". In: *Theoretical Population Biology* 23.2 (1983), pp. 202–215. ISSN: 00405809. DOI: 10.1016/0040-5809(83)90014-X. URL: http://linkinghub.elsevier.com/retrieve/pii/004058098390014X.

[83] John H Gillespie. "Is the population size of a species relevant to its evolution?" In: *Evolution* 55.11 (2001), pp. 2161–2169. ISSN: 0014-3820. DOI: 10.1111/j.0014-3820.2001.tb00732.x.

[84] John H Gillespie. "Molecular evolution over the mutational landscape". In: *Evolution* 38.5 (1984), p. 1116. ISSN: 00143820. DOI: 10.2307/2408444. URL: http://www.jstor.org/stable/2408444?origin=crossref.

[85] Tatiana Giraud, Jes S Pedersen, and Laurent Keller. "Evolution of supercolonies: The Argentine ants of southern Europe". In: *Proceedings of the National Academy of Sciences* 99.9 (2002), pp. 6075–6079. ISSN: 0027-8424. DOI: 10.1073/pnas.092694199.

[86] Nicolas Gompel and Benjamin Prud'homme. "The causes of repeated genetic evolution". In: *Developmental Biology* 332.1 (2009), pp. 36–47. ISSN: 1095-564X. DOI: 10.1016/j.ydbio.2009.04.040.

[87] Toni I Gossmann, Peter D Keightley, and Adam Eyre-Walker. "The effect of variation in the effective population size on the rate of adaptive molecular evolution in eukaryotes". In: *Genome Biology and Evolution* 4.5 (2012), pp. 658–667. ISSN: 17596653. DOI: 10.1093/gbe/evs027.

[88] Peter R Grant et al. "Convergent evolution of Darwin's finches caused by introgressive hybridization and selection". In: *Evolution; International Journal of Organic Evolution* 58.7 (2004), pp. 1588–99. ISSN: 0014-3820.

[89] Steven P Gygi et al. "Correlation between protein and mRNA abundance in yeast". In: *Molecular and Cellular Biology* 19.3 (1999), pp. 1720–1730. ISSN: 0270-7306. DOI: 10.1128/MCB.19.3.1720.

[90] Ulf Gyllensten et al. "Mitochondrial genome variation and the origin of modern humans". In: *Nature* 408.6813 (2000), pp. 708–713. ISSN: 00280836. DOI: 10.1038/35047064.

[91] Brian B Haab, Maitreya J Dunham, and Patrick O Brown. "Protein microarrays for highly parallel detection and quantitation of specific proteins and antibodies in complex solutions". In: *Genome Biology* 2.2 (2001), RESEARCH0004. ISSN: 1474-760X.

[92] John B S Haldane. "A mathematical theory of natural and artificial selection, part V: selection and mutation". In: *Mathematical Proceedings of the Cambridge Philosophical Society* 23.07 (1927), p. 838. ISSN: 0305-0041. DOI: 10.1017/S0305004100015644.

[93] Andreas Handel and Daniel E Rozen. "The impact of population size on the evolution of asexual microbes on smooth versus rugged fitness landscapes". In: *BMC Evolutionary Biology* 9 (2009), p. 236. ISSN: 1471-2148. DOI: 10.1186/1471-2148-9-236.

[94] Naoki Harada et al. "Human IgGFc binding protein (Fc $\gamma$ BP) in colonic epithelial cells exhibits mucin-like structure". In: *The Journal of Biological Chemistry* 272 (1997), pp. 15232–15241.

[95] Timothy D Harris et al. "Single-molecule DNA sequencing of a viral genome". In: *Science* 320.5872 (2008), pp. 106–9. ISSN: 1095-9203. DOI: 10.1126/science.1150427.

[96] Daniel L Hartl and Andrea G Clark. *Principles of Population Genetics*. Third. Sunderland, Massachusetts: Sinauer Associates, Inc., 1997. ISBN: 0-87893-306-9.

[97] Philip W Hedrick and Glenys Thomson. "Evidence for balancing selection at HLA". In: *Genetics* 104.3 (1983), pp. 449–456.

[98] Joachim Hermisson and Pleuni S Pennings. "Soft sweeps: molecular population genetics of adaptation from standing genetic variation". In: *Genetics* 169.4 (2005), pp. 2335–52. ISSN: 0016-6731. DOI: 10.1534/genetics.104.036947.

[99] Ruth Hershberg, Hua Tang, and Dmitri A Petrov. "Reduced selection leads to accelerated gene loss in Shigella". In: *Genome Biology* 8.8 (2007), R164. ISSN: 14656906. DOI: 10.1186/gb-2007-8-8-r164.

[100] Michael A Hollingsworth and Benjamin J Swanson. "Mucins in cancer: protection and control of the cell surface". In: *Nature Reviews. Cancer* 4.1 (2004), pp. 45–60. ISSN: 1474-175X. DOI: 10.1038/nrc1251.

[101] Rollin D Hotchkiss. "Models of genetic recombination". In: *Annual Review of Microbiology* 28 (1974), pp. 445–68. ISSN: 0066-4227. DOI: 10.1146/annurev.mi.28.100174.002305.

[102] Ronald R Hoy. "Convergent evolution of hearing". In: *Science* 338.6109 (2012), pp. 894–5. ISSN: 1095-9203. DOI: 10.1126/science.1231169.

[103] Austin L Hughes and Masatoshi Nei. "Nucleotide substitution at major histocompatibility complex class II loci: evidence for overdominant selection". In: *Proceedings of the National Academy of Sciences of the United States of America* 86 (1989), pp. 958–962. ISSN: 0027-8424. DOI: 10.1073/pnas.86.3.958.

[104] Austin L Hughes and Masatoshi Nei. "Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection". In: *Nature* 335 (1988), pp. 167–170. ISSN: 0028-0836. DOI: 10.1038/335167a0.

[105] Austin L Hughes and Meredith Yeager. "Natural selection at major histocompatibility complex loci of vertebrates". In: *Annual Review of Genetics* 32 (1998), pp. 415–435. ISSN: 0066-4197. DOI: 10.1146/annurev.genet.32.1.415.

[106] International Human Genome Sequencing Consortium. "Finishing the euchromatic sequence of the human genome". In: *Nature* 431.7011 (2004), pp. 931–945. ISSN: 0028-0836. DOI: 10.1038/nature03001.

[107] Daniel H Huson and David Bryant. "Application of phylogenetic networks in evolutionary studies". In: *Molecular biology and evolution* 23.2 (2006), pp. 254–67. ISSN: 0737-4038. DOI: 10.1093/molbev/msj030.

[108] Y. Iwasa. "Stochastic tunnels in evolutionary dynamics". In: *Genetics* 166.3 (2004), pp. 1571–1579. ISSN: 0016-6731.

[109] Jennifer A Jackson and Gerald R Fink. "Gene conversion between duplicated genetic elements in yeast". In: *Nature* 292.5821 (1981), pp. 306–311. ISSN: 0028-0836. DOI: 10.1038/292306a0.

[110] Kavita Jain, Joachim Krug, and Su-Chan Chan Park. "Evolutionary advantage of small populations on complex fitness landscapes". In: *Evolution* 65.7 (2011), pp. 1945–1955. ISSN: 00143820. DOI: 10.1111/j.1558-5646.2011.01280.x. arXiv: 1003.5380.

[111] Thomas Jarvie. "Next generation sequencing technologies". In: *Drug Discovery Today: Technologies* 2.3 (2005), pp. 255–260. ISSN: 17406749. DOI: 10.1016/j.ddtec.2005.08.003.

[112] Xiaoqian Jiang et al. “Impacts of mutation effects and population size on mutation rate in asexual populations: a simulation study”. In: *BMC Evolutionary Biology* 10 (2010), p. 298. ISSN: 1471-2148. DOI: 10.1186/1471-2148-10-298.

[113] Norman L Johnson, Samuel Kotz, and Adrienne W Kemp. *Univariate discrete distributions*. Second. New York: Wiley-Interscience; 2 edition (February 22, 1993), 1992.

[114] G Joshi-Tope et al. “Reactome: a knowledgebase of biological pathways”. In: *Nucleic acids research* 33.suppl 1 (2005), pp. D428–D432.

[115] Minoru Kanehisa et al. “KEGG as a reference resource for gene and protein annotation”. In: *Nucleic Acids Research* 44.D1 (2016), pp. D457–D462. ISSN: 0305-1048. DOI: 10.1093/nar/gkv1070.

[116] Minoru Kanehisa et al. “KEGG: new perspectives on genomes, pathways, diseases and drugs”. In: *Nucleic Acids Research* 45.D1 (2017), pp. D353–D361. ISSN: 0305-1048. DOI: 10.1093/nar/gkw1092.

[117] Talia Karasov, Philipp W. Messer, and Dmitri A. Petrov. “Evidence that adaptation in Drosophila is not limited by mutation at single sites”. In: *PLoS Genetics* 6.6 (2010). Ed. by Harmit S. Malik, e1000924. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1000924.

[118] Arek Kasprzyk. “BioMart: driving a paradigm change in biological data management”. In: *Database : the Journal of Biological Databases and Curation* 2011 (2011), bar049. ISSN: 1758-0463. DOI: 10.1093/database/bar049.

[119] Minae Kawashima et al. “Evolutionary analysis of classical HLA class I and II genes suggests that recent positive selection acted on DPB1*04:01 in Japanese population”. In: *PloS One* 7.10 (2012), e46806. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0046806.

[120] Anthony D Keefe, Supriya Pai, and Andrew Ellington. “Aptamers as therapeutics”. In: *Nature Reviews Drug Discovery* 9.7 (2010), pp. 537–550. ISSN: 1474-1776. DOI: 10.1038/nrd3141.

[121] Peter D Keightley and Adam Eyre-Walker. “Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies”. In: *Genetics* 177.4 (2007), pp. 2251–2261. ISSN: 0016-6731. DOI: 10.1534/genetics.107.080663.

[122] Fahad Khalid et al. "Genonets server-a web server for the construction, analysis and visualization of genotype networks". In: *Nucleic Acids Research* 44.W1 (2016), gkw313. ISSN: 1362-4962. DOI: 10.1093/nar/gkw313.

[123] Yuseob Kim and H Allen Orr. "Adaptation in sexuals vs. asexuals: clonal interference and the Fisher-Muller model". eng. In: *Genetics* 171.3 (2005), pp. 1377–1386. DOI: genetics.105.045252[pii]10.1534/genetics.105.045252. URL: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve{\&}db=PubMed{\&}dopt=Citation{\&}list{_}uids=16020775.

[124] Motoo Kimura. "On the probability of fixation of mutant genes in a population". In: *Genetics* 47.6 (1962), pp. 713–719.

[125] Motoo Kimura. "Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution". In: *Nature* 267 (1977), pp. 275–276. ISSN: 0028-0836. DOI: 10.1038/267275a0.

[126] Motoo Kimura and Tomoko Ohta. "Protein polymorphism as a phase of molecular evolution". In: *Nature* 229.5285 (1971), pp. 467–469.

[127] Taishin Kin et al. "fRNAdb: A platform for mining/annotating functional RNA candidates from non-coding RNA sequences". In: *Nucleic Acids Research* 35.SUPPL. 1 (2007). ISSN: 03051048. DOI: 10.1093/nar/gkl837.

[128] Natalia L Komarova, Anirvan Sengupta, and Martin A Nowak. "Mutation–selection networks of cancer initiation: tumor suppressor genes and chromosomal instability". In: *Journal of Theoretical Biology* 223.4 (2003), pp. 433–450. ISSN: 00225193. DOI: 10.1016/S0022-5193(03)00120-6.

[129] Augustine Kong et al. "Fine-scale recombination rate differences between sexes, populations and individuals". In: *Nature* 467.7319 (2010), pp. 1099–1103. ISSN: 0028-0836. DOI: 10.1038/nature09525.

[130] Eugene V Koonin. "Evolution of genome architecture". In: *The International Journal of Biochemistry & Cell Biology* 41.2 (2009), pp. 298–306. ISSN: 13572725. DOI: 10.1016/j.biocel.2008.09.015.

[131] Jeffrey M Koshi and Richard A Goldstein. "Probabilistic reconstruction of ancestral protein sequences". In: *Journal of Molecular Evolution* 42.2 (1996), pp. 313–320. ISSN: 00222844. DOI: 10.1007/BF02198858.

[132] Carolin Kosiol et al. "Patterns of positive selection in six mammalian genomes". In: *PLoS Genetics* 4.8 (2008), e1000144. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1000144.

[133] Karin Kriener et al. "Convergent evolution of major histocompatibility complex molecules in humans and New World monkeys". In: *Immunogenetics* 51 (2000), pp. 169–178. ISSN: 0093-7711. DOI: 10.1007/s002510050028.

[134] Sergey Kryazhimskiy and Joshua B Plotkin. "The population genetics of dN/dS". In: *PLoS Genetics* 4.12 (2008), e1000304. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1000304.

[135] Sergey Kryazhimskiy, Daniel P Rice, and Michael M Desai. "Population subdivision and adaptation in asexual populations of *Saccharomyces cerevisiae*". In: *Evolution* 66.6 (2012), pp. 1931–1941. ISSN: 00143820. DOI: 10.1111/j.1558-5646.2011.01569.x.

[136] Victor Kunin et al. "The net of life: Reconstructing the microbial phylogenetic network". In: *Genome Research* 15.7 (2005), pp. 954–959. ISSN: 1088-9051. DOI: 10.1101/gr.3666505.

[137] Josianne Lachapelle, Joshua Reid, and Nick Colegrave. "Repeatability of adaptation in experimental populations of different sizes". In: *Proceedings of the Royal Society B: Biological Sciences* 282.1805 (2015), pp. 20143033–20143033. ISSN: 0962-8452. DOI: 10.1098/rspb.2014.3033.

[138] E S Lander et al. "Initial sequencing and analysis of the human genome". In: *Nature* 409.6822 (2001), pp. 860–921. ISSN: 0028-0836. DOI: 10.1038/35057062.

[139] Robert Lanfear, Hanna Kokko, and Adam Eyre-Walker. "Population size and the rate of evolution". In: *Trends in Ecology & Evolution* 29.1 (2014), pp. 33–41. ISSN: 01695347. DOI: 10.1016/j.tree.2013.09.009.

[140] David S Latchman. "Transcription factors: an overview". In: *The International Journal of Biochemistry & Cell Biology* 29.12 (1997), pp. 1305–1312. ISSN: 13572725. DOI: 10.1016/S1357-2725(97)00085-X.

[141] Rechard E Lenski et al. "Long-term experimental evolution in *Escherichia coli*. I. Adaptation and divergence during 2,000 generations". In: *American Naturalist* 138.6 (1991), pp. 1315–1341.

[142] Richard E Lenski and Michael Travisano. "Dynamics of adaptation and diversification: a 10,000-generation experiment with bacterial populations". In: *Proceedings of the National Academy of Sciences of the United States of America* 91.15 (1994), pp. 6808–6814. ISSN: 0027-8424. DOI: 10.1073/pnas.91.15.6808.

[143] Tobias L Lenz. "Computational prediction of MHC II-antigen binding supports divergent allele advantage and explains strans-species polymorphism". In: *Evolution* 65.8 (2011), pp. 2380–2390. ISSN: 00143820. DOI: 10.1111/j.1558-5646.2011.01288.x.

[144] Xiao-yong Li et al. "Transcription factors bind thousands of active and inactive regions in the Drosophila blastoderm". In: *PLoS Biology* 6.2 (2008), e27. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.0060027.

[145] David J Lipman and W John Wilbur. "Modeling neutral and selective evolution of protein folding". In: *Proceedings of the Royal Society B Biological Sciences* 245.1312 (1991), pp. 7–11. ISSN: 0962-8452. DOI: 10.1098/rspb.1991.0081.

[146] David J Lipman and W John Wilbur. "Modelling neutral and selective evolution of protein folding". In: *Proceedings. Biological sciences* 245.1312 (1991), pp. 7–11. ISSN: 0962-8452. DOI: 10.1098/rspb.1991.0081.

[147] Yang Liu et al. "Convergent sequence evolution between echolocating bats and dolphins". In: *Current Biology* 20.2 (2010), R53–R54. ISSN: 09609822. DOI: 10.1016/j.cub.2009.11.058.

[148] Alexander E Lobkovsky, Yuri I Wolf, and Eugene V Koonin. "Predictability of evolutionary trajectories in fitness landscapes". In: *PLoS Computational Biology* 7.12 (2011), e1002302. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1002302.

[149] Laurence Loewe and Brian Charlesworth. "Inferring the distribution of mutational effects on fitness in Drosophila". In: *Biology Letters* 2.3 (2006), pp. 426–430. ISSN: 1744-9561. DOI: 10.1098/rsbl.2006.0481.

[150] Ronny Lorenz et al. "ViennaRNA Package 2.0". In: *Algorithms for Molecular Biology* 6.1 (2011), p. 26. ISSN: 1748-7188. DOI: 10.1186/1748-7188-6-26.

[151] Jonathan B Losos. "Convergence, adaptation, and constraint". In: *Evolution* 65.7 (2011), pp. 1827–1840. ISSN: 00143820. DOI: 10.1111/j.1558-5646.2011.01289.x.

[152] Jonathan B Losos et al. *The Princeton Guide to Evolution*. Princeton: Princeton University Press, 2013. ISBN: 0691149771.

[153] João M Lourenço, Sylvain Glémin, and Nicolas Galtier. "The rate of molecular adaptation in a changing environment". In: *Molecular Biology and Evolution* 30.6 (2013), pp. 1292–1301. ISSN: 07374038. DOI: 10.1093/molbev/mst026.

[154] Gordon Luikart et al. "Estimation of census and effective population sizes: the increasing usefulness of DNA-based approaches". In: *Conservation Genetics* 11.2 (2010), pp. 355–373. ISSN: 1566-0621. DOI: 10.1007/s10592-010-0050-7.

[155] Ante S Lundberg and Hugh O McDevitt. "Evolution of major histocompatibility complex class II allelic diversity: direct descent in mice and humans". In: *Proceedings of the National Academy of Sciences of the United States of America* 89.July (1992), pp. 6545–6549. ISSN: 0027-8424. DOI: 10.1073/pnas.89.14.6545.

[156] Michael Lynch and John S Conery. "The origins of genome complexity". In: *Science* 302.5649 (2003), pp. 1401–1404. ISSN: 1095-9203. DOI: 10.1126/science.1089370.

[157] Michael Lynch and Russell Lande. "The critical effective size for a genetically secure population". In: *Animal Conservation* 01.01 (1998), S136794309822110X. ISSN: 13679430. DOI: 10.1017/S136794309822110X.

[158] Nicole Maca-Meyer et al. "Major genomic mitochondrial lineages delineate early human expansions". In: *BMC Genetics* 2.1 (2001), p. 13. ISSN: 14712156. DOI: 10.1186/1471-2156-2-13.

[159] M Mann. "Quantitative proteomics?" In: *Nature Biotechnology* 17.10 (1999), pp. 954–5. ISSN: 1087-0156. DOI: 10.1038/13646.

[160] Joyce M Manzella et al. "Insulin induction of ornithine decarboxylase. Importance of mRNA secondary structure and phosphorylation of eucaryotic initiation factors eIF-4B and eIF-4E". In: *The Journal of Biological Chemistry* 266.4 (1991), pp. 2383–9. ISSN: 0021-9258.

[161] Marcel Margulies et al. "Genome sequencing in microfabricated high-density picolitre reactors". In: *Nature* 437.7057 (2005), pp. 376–80. ISSN: 1476-4687. DOI: 10.1038/nature03959.

[162] Nicholas R Markham and Michael Zuker. "DINAMelt web server for nucleic acid melting prediction". In: *Nucleic Acids Research* 33.Web Server issue (2005), W577–81. ISSN: 1362-4962. DOI: 10.1093/nar/gki591.

[163] David H Mathews et al. "Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure". In: *Proceedings of the National Academy of Sciences of the United States of America* 101.19 (2004), pp. 7287–92. ISSN: 0027-8424. DOI: 10.1073/pnas.0401799101.

[164] Joao F Matias Rodrigues and Andreas Wagner. "Evolutionary plasticity and innovations in complex metabolic reaction networks". In: *PLoS Computational Biology BIOLOGY* 5.12 (2009), e1000613. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1000613.

[165] John Maynard Smith. "Natural selection and the concept of a protein space". In: *Nature* 225.5232 (1970), pp. 563–564. ISSN: 0028-0836. DOI: 10.1038/225563a0.

[166] Bruce A McDonald and Celeste Linde. "Pathogen population genetics, evolutionary potential, and durable resistance". In: *Annual Review of Phytopathology* 40 (2002), pp. 349–79. ISSN: 0066-4286. DOI: 10.1146/annurev.phyto.40.120501.101443.

[167] Glenn H McGall and Fred C Christians. "High-density genechip oligonucleotide probe arrays". In: *Advances in Biochemical Engineering/Biotechnology* 77 (2002), pp. 21–42. ISSN: 0724-6145.

[168] Gil A. McVean et al. "An integrated map of genetic variation from 1,092 human genomes". In: *Nature* 491.7422 (2012), pp. 56–65. ISSN: 0028-0836. DOI: 10.1038/nature11632.

[169] Gilean A T McVean et al. "The fine-scale structure of recombination rate variation in the human genome". In: *Science* 304.5670 (2004), pp. 581–4. ISSN: 1095-9203. DOI: 10.1126/science.1092500.

[170] R Milo et al. "Network motifs: simple building blocks of complex networks". In: *Science* 298.5594 (2002), pp. 824–7. ISSN: 1095-9203. DOI: 10.1126/science.298.5594.824.

[171] Jeffrey C Mogul. "Emergent (mis)behavior vs. complex software systems". In: *ACM SIGOPS Operating Systems Review* 40.4 (2006), p. 293. ISSN: 01635980. DOI: 10.1145/1218063.1217964.

[172] Nancy A Moran, John P McCutcheon, and Atsushi Nakabachi. "Genomics and evolution of heritable bacterial symbionts". In: *Annual Review of Genetics* 42.1 (2008), pp. 165–190. ISSN: 0066-4197. DOI: 10.1146/annurev.genet.41.110306.130119.

[173] Bernard M E Moret et al. "Phylogenetic networks: Modeling, reconstructibility, and accuracy". In: *IEEE-ACM Transactions on Computational Biology and Bioinformatics* 1.1 (2004), pp. 13–23. ISSN: 1545-5963. DOI: 10.1109/TCBB.2004.10.

[174] David A Morrison. "Networks in phylogenetic analysis: new tools for population biology". In: *International Journal for Parasitology* 35.5 (2005), pp. 567–582. ISSN: 0020-7519. DOI: 10.1016/j.ijpara.2005.02.007.

[175] Stefanie A Mortimer, Mary Anne Kidwell, and Jennifer A Doudna. "Insights into RNA structure and function from genome-wide studies". In: *Nature Reviews Genetics* 15.7 (2014), pp. 469–479. ISSN: 1471-0056. DOI: 10.1038/nrg3681.

[176] Sonali Mukherjee et al. "Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays". In: *Nature Genetics* 36.12 (2004), pp. 1331–9. ISSN: 1061-4036. DOI: 10.1038/ng1473.

[177] Hermann J Muller. "Some genetic aspects of sex". In: *The American Naturalist* 66.703 (1932), pp. 118–138. ISSN: 0003-0147. DOI: 10.1086/280418.

[178] Ville Mustonen and Michael Lässig. "Evolutionary population genetics of promoters: predicting binding sites and functional phylogenies". In: *Proceedings of the National Academy of Sciences of the United States of America* 102.44 (2005), pp. 15936–15941. ISSN: 0027-8424. DOI: 10.1073/pnas.0505537102.

[179] Ville Mustonen and Michael Lässig. "From fitness landscapes to seascapes: non-equilibrium dynamics of selection and adaptation". In: *Trends in Genetics : TIG* 25.3 (2009), pp. 111–9. ISSN: 0168-9525. DOI: 10.1016/j.tig.2009.01.002.

[180] National Center for Biotechnology Information. *Complete genomes: Viruses*. URL: https://www.ncbi.nlm.nih.gov/genomes/GenomesGroup.cgi?taxid=10239.

[181] Masatoshi Nei and Sudhir Kumar. *Molecular Evolution and Phylogenetics*. 1st edition. Oxford University Press, 2000.

[182] M E J Newman. "Assortative mixing in networks". In: *Physical Review Letters* 89.20 (2002), p. 208701. ISSN: 0031-9007. DOI: 10.1103/PhysRevLett.89.208701.

[183] M E J Newman. "Modularity and community structure in networks". In: *Proceedings of the National Academy of Sciences of the United States of America* 103.23 (2006), pp. 8577–82. ISSN: 0027-8424. DOI: 10.1073/pnas.0601602103.

[184] Rasmus Nielsen and Ziheng Yang. "Estimating the distribution of selection coefficients from phylogenetic data with applications to mitochondrial and viral DNA". In: *Molecular Biology and Evolution* 20.8 (2003), pp. 1231–1239. ISSN: 0737-4038. DOI: 10.1093/molbev/msg147.

[185] Rasmus Nielsen et al. "A scan for positively selected genes in the genomes of humans and chimpanzees". In: *PLoS Biology* 3.6 (2005), e170. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.0030170.

[186] Paul J Norman et al. "Co-evolution of human leukocyte antigen (HLA) class I ligands with killer-cell immunoglobulin-like receptors (KIR) in a genetically diverse population of sub-Saharan Africans". In: *PLoS Genetics* 9.10 (2013), e1003938. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1003938.

[187] Ruth Nussinov et al. "Algorithms for loop matchings". In: *SIAM Journal on Applied Mathematics* 35.1 (1978), pp. 68–82. DOI: 10.1137/0135006.

[188] Howard Ochman, Jeffrey G Lawrence, and Eduardo A Groisman. "Lateral gene transfer and the nature of bacterial innovation". In: *Nature* 405.6784 (2000), pp. 299–304. ISSN: 0028-0836. DOI: 10.1038/35012500.

[189] Tomoko Ohta. "Role of diversifying selection and gene conversion in evolution of major histocompatibility complex loci". In: *Proceedings of the National Academy of Sciences of the United States of America* 88 (1991), pp. 6716–6720. ISSN: 0027-8424. DOI: 10.1073/pnas.88.15.6716.

[190] Tomoko Ohta. "The nearly neutral theory Of molecular evolution". In: *Annual Review of Ecology and Systematics* 23.1992 (1992), pp. 263–286. ISSN: 00664162. DOI: 10.2307/2097289.

[191] Tomoko Ohta and John H Gillespie. "Development of neutral and nearly neutral theories". In: *Theoretical Population Biology* 49.2 (1996), pp. 128–42. ISSN: 1096-0325. DOI: 10.1006/tpbi.1996.0007.

[192] Taras K Oleksyk, Michael W Smith, and Stephen J O'Brien. "Genome-wide scans for footprints of natural selection". In: *Philosophical Transactions of the Royal Society of London. Series B, Biological sciences* 365.1537 (2010), pp. 185–205. ISSN: 1471-2970. DOI: 10.1098/rstb.2009.0219.

[193] J-P Onnela et al. "Structure and tie strengths in mobile communication networks". In: *Proceedings of the National Academy of Sciences of the United States of America* 104.18 (2007), pp. 7332–6. ISSN: 0027-8424. DOI: 10.1073/pnas.0610245104.

[194] Allen H Orr. "The rate of adaptation in asexuals". In: *Genetics* 155.2 (2000), pp. 961–968. ISSN: 00166731.

[195] Sarah P Otto and Aleeza C Gerstein. "Why have sex? The population genetics of sex and recombination". In: *Biochemical Society Transactions* 34.Pt 4 (2006), pp. 519–22. ISSN: 0300-5127. DOI: 10.1042/BST0340519.

[196] Oxford Nanopore Technologies. *Nanopore*. URL: https://nanoporetech.com/.

[197] Pacific Biosciences, Menlo Park, USA. *PacBio*. URL: http://www.pacb.com/.

[198] Su-Chan Park and Joachim Krug. "Clonal interference in large populations". In: *Proceedings of the National Academy of Sciences of the United States of America* 104.46 (2007), pp. 18135–40. ISSN: 1091-6490. DOI: 10.1073/pnas.0705778104. arXiv: arXiv:0711.1989.

[199] Joe Parker et al. "Genome-wide signatures of convergent evolution in echolocating mammals". In: *Nature* 502.7470 (2013), pp. 228–31. ISSN: 1476-4687. DOI: 10.1038/nature12511.

[200] Michael S Paul and Brenda L Bass. "Inosine exists in mRNA at tissue-specific levels and is most abundant in brain mRNA". In: *The EMBO Journal* 17.4 (1998), pp. 1120–1127. ISSN: 14602075. DOI: 10.1093/emboj/17.4.1120.

[201] Joshua L Payne and Andreas Wagner. "The robustness and evolvability of transcription factor binding sites". In: *Science* 343.6173 (2014), pp. 875–877. ISSN: 0036-8075. DOI: 10.1126/science.1249046.

[202] Pleuni S Pennings and Joachim Hermisson. "Soft sweeps II–molecular population genetics of adaptation from recurrent mutation or migration". In: *Molecular Biology and Evolution* 23.5 (2006), pp. 1076–84. ISSN: 0737-4038. DOI: 10.1093/molbev/msj117.

[203] Lilia Perfeito et al. "Adaptive mutations in bacteria: high rate and small effects". In: *Science* 317.5839 (2007), pp. 813–815. ISSN: 0036-8075. DOI: 10.1126/science.1142284. URL: http://www.sciencemag.org/cgi/doi/10.1126/science.1142284.

[204] Ernesto Picardi, ed. *RNA bioinformatics (methods in molecular biology)*. Humana Press, 2015. ISBN: 978-1-4939-2291-8.

[205] Frank J Poelwijk et al. "Empirical fitness landscapes reveal accessible evolutionary paths". In: *Nature* 445.7126 (2007), pp. 383–386. ISSN: 1476-4687. DOI: 10.1038/nature05451.

[206] Cornelia Pokalyuk et al. "Competing islands limit the rate of adaptation in structured populations". In: *Theoretical Population Biology* 90 (2013), pp. 1–11. ISSN: 00405809. DOI: 10.1016/j.tpb.2013.08.001.

[207] Elena A. Ponomarenko et al. "The size of the human proteome: the width and depth". In: *International Journal of Analytical Chemistry* 2016 (2016), pp. 1–6. ISSN: 1687-8760. DOI: 10.1155/2016/7436849.

[208] Nathan D Price, Jennifer L Reed, and Bernhard Ø Palsson. "Genome-scale models of microbial cells: evaluating the consequences of constraints". In: *Nature Reviews. Microbiology* 2.11 (2004), pp. 886–97. ISSN: 1740-1526. DOI: 10.1038/nrmicro1023.

[209] Estelle Proux et al. "Selectome: a database of positive selection". In: *Nucleic Acids Research* 37.Database issue (2009), pp. D404–7. ISSN: 1362-4962. DOI: 10.1093/nar/gkn768.

[210] Marc Pybus et al. "1000 Genomes Selection Browser 1.0: a genome browser dedicated to signatures of natural selection in modern humans". In: *Nucleic Acids Research* 42.1 (2014), pp. D903–9. ISSN: 1362-4962. DOI: 10.1093/nar/gkt1188.

[211] Ali R. Vahdati and Andreas Wagner. "Parallel or convergent evolution in human population genomic data revealed by genotype networks". In: *BMC Evolutionary Biology* 16.1 (2016), p. 154. ISSN: 1471-2148. DOI: 10.1186/s12862-016-0722-0.

[212] Jang B. Rampal, ed. *Microarrays*. Totowa, NJ: Humana Press, 2007. ISBN: 978-1-58829-589-7. DOI: 10.1007/978-1-59745-303-5.

[213] Neil D Rawlings and Alan J Barrett. "Evolutionary families of peptidases". In: *Biochemical Journal* 290.1 (1993), pp. 205–218. ISSN: 0264-6021. DOI: 10.1042/bj2900205.

[214] Jüri Reimand, Tambet Arak, and Jaak Vilo. "g:Profiler–a web server for functional interpretation of gene lists (2011 update)". In: *Nucleic Acids Research* 39.Web Server issue (2011), W307–15. ISSN: 1362-4962. DOI: 10.1093/nar/gkr378.

[215] Jessica S Reuter and David H Mathews. "RNAstructure: software for RNA secondary structure prediction and analysis". In: *BMC Bioinformatics* 11.1 (2010), p. 129. ISSN: 1471-2105. DOI: 10.1186/1471-2105-11-129.

[216] Daniel E Rozen et al. "Heterogeneous adaptive trajectories of small populations on complex fitness landscapes". In: *PloS One* 3.3 (2008), e1715. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0001715.

[217] Pardis C Sabeti et al. "Detecting recent positive selection in the human genome from haplotype structure". In: *Nature* 419.6909 (2002), pp. 832–837. ISSN: 0028-0836. DOI: 10.1038/nature01140.

[218] Yasubumi Sakakibara et al. "Stochastic context-free grammers for tRNA modeling". In: *Nucleic Acids Research* 22.23 (1994), pp. 5112–5120. ISSN: 0305-1048. DOI: 10.1093/nar/22.23.5112.

[219] Meena Kishore Sakharkar, Vincent T K Chow, and Pandjassarame Kangueane. "Distributions of exons and introns in the human genome". In: *In Silico Biology* 4.4 (2004), pp. 387–393. ISSN: 1386-6338. DOI: 2004040032[pii].

[220] F Sanger, S Nicklen, and A R Coulson. "DNA sequencing with chain-terminating inhibitors". In: *Proceedings of the National Academy of Sciences of the United States of America* 74.12 (1977), pp. 5463–7. ISSN: 0027-8424.

[221] Stanley A Sawyer. "GENECONV: A computer package for the statistical detection of gene conversion". In: *Distributed by the author, Department of Mathematics, Washington University in St. Louis* (1999).

[222] Mark Schena et al. "Quantitative monitoring of gene expression patterns with a complementary DNA microarray". In: *Science* 270.5235 (1995), pp. 467–70. ISSN: 0036-8075.

[223] Peter Schuster. "Prediction of RNA secondary structures: from theory to models and real molecules". In: *Reports on Progress in Physics* 69.5 (2006), pp. 1419–1477. ISSN: 0034-4885. DOI: 10.1088/0034-4885/69/5/R04.

[224] Peter Schuster and Walter Fontana. "Chance and necessity in evolution: lessons from RNA". In: *Physica D: Nonlinear Phenomena* 133.1-4 (1999), pp. 427–452. ISSN: 01672789. DOI: 10.1016/S0167-2789(99)00076-7.

[225] Peter Schuster et al. "From sequences to shapes and back - a case study in RNA secondary structures". In: *Proceedings of the Royal Society B Biological Sciences* 255.1344 (1994), pp. 279–284. ISSN: 0962-8452. DOI: 10.1098/rspb.1994.0040.

[226] Kurt Schwenk. "A utilitarian approach to evolutionary constraint". In: *Zoology* 98 (1995), pp. 251–262.

[227] Jay Shendure and Hanlee Ji. "Next-generation DNA sequencing". In: *Nature Biotechnology* 26.10 (2008), pp. 1135–1145. ISSN: 1087-0156. DOI: 10.1038/nbt1486.

[228] Jay Shendure et al. "Accurate multiplex polony sequencing of an evolved bacterial genome". In: *Science* 309.5741 (2005), pp. 1728–32. ISSN: 1095-9203. DOI: 10.1126/science.1117389.

[229] Olin K Silander, Olivier Tenaillon, and Lin Chao. "Understanding the evolutionary fate of finite populations: the dynamics of mutational effects". In: *PLoS Biology* 5.4 (2007), e94. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.0050094.

[230] Sergey V Solomatin et al. "Multiple native states reveal persistent ruggedness of an RNA folding landscape". In: *Nature* 463.7281 (2010), pp. 681–684. ISSN: 0028-0836. DOI: 10.1038/nature08717.

[231] Edwin Southern, Kalim Mir, and Mikhail Shchepinov. "Molecular interactions on microarrays". In: *Nature Genetics* 21 (1999), pp. 5–9. ISSN: 10614036. DOI: 10.1038/4429.

[232] Olaf Sporns, Giulio Tononi, and Rolf Kötter. "The Human Connectome: A structural description of the human brain". In: *PLoS Computational Biology* 1.4 (2005), e42. ISSN: 1553-734X. DOI: 10.1371/journal.pcbi.0010042.

[233] Jason E Stajich and Matthew W Hahn. "Disentangling the effects of demography and selection in human history". In: *Molecular Biology and Evolution* 22.1 (2005), pp. 63–73. ISSN: 0737-4038. DOI: 10.1093/molbev/msh252.

[234] Craig A Stockwell, Andrew P Hendry, and Michael T Kinnison. "Contemporary evolution meets conservation biology". In: *Trends in Ecology & Evolution* 18.2 (2003), pp. 94–101. ISSN: 01695347. DOI: 10.1016/S0169-5347(02)00044-7.

[235] Ivan G Szendro et al. "Predictability of evolution depends nonmonotonically on population size". In: *Proceedings of the National Academy of Sciences* 110.2 (2013), pp. 571–576. ISSN: 0027-8424. DOI: 10.1073/pnas.1213613110.

[236] Damian Szklarczyk et al. "STRING v10: protein-protein interaction networks, integrated over the tree of life". In: *Nucleic Acids Research* 43.D1 (2015), pp. D447–D452. ISSN: 0305-1048. DOI: 10.1093/nar/gku1003.

[237] Fumio Tajima. "Statistical method for testing the neutral mutation hypothesis by DNA polymorphism". In: *Genetics* 123.3 (1989), pp. 585–595. ISSN: 0016-6731.

[238] Naoyuki Takahata. "Allelic genealogy and human evolution". In: *Molecular Biology and Evolution* 10.1 (1993), pp. 2–22. ISSN: 0737-4038.

[239] Asif U Tamuri, Mario dos Reis, and Richard A Goldstein. "Estimating the distribution of selection coefficients from phylogenetic data using sitewise mutation-selection models". In: *Genetics* 190.3 (2012), pp. 1101–1115. ISSN: 0016-6731. DOI: 10.1534/genetics.111.136432.

[240] Elliot Tan et al. "A four-way junction accelerates hairpin ribozyme folding via a discrete intermediate". In: *Proceedings of the National Academy of Sciences* 100.16 (2003), pp. 9308–9313. ISSN: 0027-8424. DOI: 10.1073/pnas.1233536100.

[241] *The CAIDA UCSD*. URL: http://www.caida.org.

[242] Ignacio Tinoco and Carlos Bustamante. "How RNA folds". In: *Journal of Molecular Biology* 293.2 (1999), pp. 271–81. ISSN: 0022-2836. DOI: 10.1006/jmbi.1999.3001.

[243] Douglas H Turner and David H Mathews, eds. *RNA structure determination: methods and protocols (methods in molecular biology)*. 1st editio. Humana Press, 2016. ISBN: 978-1-4939-6433-8.

[244] Eray Tuzun et al. "Fine-scale structural variation of the human genome". In: *Nature Genetics* 37.7 (2005), pp. 727–732. ISSN: 1061-4036. DOI: 10.1038/ng1562.

[245] UniProt Consortium. "UniProt: a hub for protein information". In: *Nucleic Acids Research* 43.Database issue (2015), pp. D204–12. ISSN: 1362-4962. DOI: 10.1093/nar/gku989.

[246] Alexandra I Vatsiou, Eric Bazin, and Oscar E Gaggiotti. "Detection of selective sweeps in structured populations: a comparison of recent methods". en. In: *Molecular Ecology* 25.1 (2016), pp. 89–103. ISSN: 09621083. DOI: 10.1111/mec.13360.

[247] J A G M de Visser and D E Rozen. "Limits to adaptation in asexual populations". In: *Journal of Evolutionary Biology* 18.4 (2005), pp. 779–88. ISSN: 1010-061X. DOI: 10.1111/j.1420-9101.2005.00879.x.

[248] J Arjan G M de Visser, Tim F Cooper, and Santiago F Elena. "The causes of epistasis". In: *Proceedings. Biological sciences / The Royal Society* 278.1725 (2011), pp. 3617–24. ISSN: 1471-2954. DOI: 10.1098/rspb.2011.1537.

[249] J Arjan G M de Visser and Joachim Krug. "Empirical fitness landscapes and the predictability of evolution". In: *Nature Reviews. Genetics* 15.7 (2014), pp. 480–90. ISSN: 1471-0064. DOI: 10.1038/nrg3744.

[250] Edward J Vowles and William Amos. "Evidence for widespread convergent evolution around human microsatellites". In: *PLoS Biology* 2.8 (2004). Ed. by David Penny, e199. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.0020199.

[251] Andreas Wagner. "A genotype network reveals homoplastic cycles of convergent evolution in influenza A (H3N2) haemagglutinin". In: *Proceedings of the Royal Society B: Biological Sciences* 281.1786 (2014), p. 20132763. ISSN: 0962-8452. DOI: 10.1098/rspb.2013.2763.

[252] Andreas Wagner. "Genotype networks shed light on evolutionary constraints". In: *Trends in Ecology & Evolution* 26.11 (2011), pp. 577–84. ISSN: 0169-5347. DOI: 10.1016/j.tree.2011.07.001.

[253] Andreas Wagner. "Mutational robustness accelerates the origin of novel RNA phenotypes through phenotypic plasticity". In: *Biophysical journal* 106.4 (2014), pp. 955–65. ISSN: 1542-0086. DOI: 10.1016/j.bpj.2014.01.003.

[254] Andreas Wagner. "Neutralism and selectionism: a network-based reconciliation". In: *Nature Reviews. Genetics* 9.12 (2008), pp. 965–74. ISSN: 1471-0064. DOI: 10.1038/nrg2473.

[255] Andreas Wagner. *The origins of evolutionary innovations: A theory of transformative change in living systems*. New York, NY, USA: Oxford University Press, 2011. ISBN: 978-0199692606.

[256] David B Wake. "Homoplasy: the result of natural selection, or evidence of design limitations?" In: *American Naturalist* 138.3 (1991), pp. 543–567.

[257] David B Wake, Marvalee H Wake, and Chelsea D Specht. "Homoplasy: from detecting pattern to determining process and mechanism of evolution". In: *Science* 331.6020 (2011), pp. 1032–5. ISSN: 1095-9203. DOI: 10.1126/science.1188545.

[258] Christina Waldsich, ed. *RNA Folding: Methods and Protocols (Methods in Molecular Biology)*. 1st editio. Humana Press, 2014. ISBN: 978-1-62703-667-2.

[259] Daniel M Weinreich and Lin Chao. "Rapid evolutionary escape by large populations from local fitness peaks is likely in nature". In: *Evolution; International Journal of Organic Evolution* 59.6 (2005), pp. 1175–1182. ISSN: 0014-3820. DOI: 10.1111/j.0014-3820.2005.tb01769.x.

[260] Matthew T Weirauch et al. "Determination and inference of eukaryotic transcription factor sequence specificity". In: *Cell* 158.6 (2014), pp. 1431–1443. ISSN: 00928674. DOI: 10.1016/j.cell.2014.08.009.

[261] Matthew T Weirauch et al. "Determination and inference of eukaryotic transcription factor sequence specificity". In: *Cell* 158.6 (2014), pp. 1431–43. ISSN: 1097-4172. DOI: 10.1016/j.cell.2014.08.009.

[262] Daniel B Weissman, Marcus W Feldman, and Daniel S Fisher. "The rate of fitness-valley crossing in sexual populations". In: *Genetics* 186 (2010), pp. 1389–1410. DOI: 10.1534/genetics.110.123240.

[263] Claus O. Wilke. "The speed of adaptation in large asexual populations". In: *Genetics* 167.August (2004), pp. 2045–2053. ISSN: 0016-6731. DOI: 10.1534/genetics.104.027136. arXiv: 0402009 [q-bio].

[264] Michael J Wiser, Noah Ribeck, and Richard E Lenski. "Long-term dynamics of adaptation in asexual populations". In: *Science* 342.6164 (2013), pp. 1364–1367. ISSN: 0036-8075. DOI: 10.1126/science.1243357.

[265] Alex Wong and Rees Kassen. "Parallel evolution and local differentiation in quinolone resistance in *Pseudomonas aeruginosa*". In: *Microbiology* 157.Pt 4 (2011), pp. 937–44. ISSN: 1465-2080. DOI: 10.1099/mic.0.046870-0.

[266] Alex Wong and Kimberley Seguin. "Effects of genotype on rates of substitution during experimental evolution". In: *Evolution* 69.7 (2015), pp. 1772–1785. ISSN: 15585646. DOI: 10.1111/evo.12700.

[267] Megan Woolfit. "Effective population size and the rate and pattern of nucleotide substitutions". In: *Biology Letters* 5.April (2009), pp. 417–420. ISSN: 1744-9561. DOI: 10.1098/rsbl.2009.0155.

[268] Megan Woolfit and Lindell Bromham. "Increased rates of sequence evolution in endosymbiotic bacteria and fungi with small effective population sizes". In: *Molecular Biology and Evolution* 20.9 (2003), pp. 1545–1555. ISSN: 0737-4038. DOI: 10.1093/molbev/msg167.

[269] Megan Woolfit and Lindell Bromham. "Population size and molecular evolution on islands". In: *Proceedings of the Royal Society B: Biological Sciences* 272.1578 (2005), pp. 2277–2282. ISSN: 0962-8452. DOI: 10.1098/rspb.2005.3217.

[270] Sewall Wright. "Evolution in Mendelian Populations". In: *Genetics* 16.2 (1931), pp. 97–159. ISSN: 0016-6731.

[271] Sewall Wright. *The roles of mutation, inbreeding, crossbreeding and selection in evolution*. 1932. DOI: citeulike-article-id:1586133.

[272] Catherine J Wu, ed. *Protein microarray for disease analysis*. Humana Press, 2011. DOI: 10.1007/978-1-61779-043-0. URL: https://doi.org/10.1007{\%}2F978-1-61779-043-0.

[273] Stefan Wuchty et al. "Complete suboptimal folding of RNA and the stability of secondary structures". In: *Biopolymers* 49.2 (1999), pp. 145–165. ISSN: 0006-3525. DOI: 10.1002/(SICI)1097-0282(199902)49:2<145::AID-BIP4>3.0.CO;2-G.

[274] Ziheng Yang and Joseph P Bielawski. "Statistical methods for detecting molecular adaptation". In: *Trends in Ecology & Evolution* 15.12 (2000), pp. 496–503. ISSN: 01695347. DOI: 10.1016/S0169-5347(00)01994-7. arXiv: S0169-5347(00)01994-7.

[275] Ziheng Yang, Sudhir Kumar, and Masatoshi Nei. "A new method of inference of ancestral nucleotide and amino acid sequences". In: *Genetics* 141.4 (1995), pp. 1641–1650. ISSN: 00166731. DOI: 8601501.

[276] Meredith Yeager, Sudhir Kumar, and Austin L Hughes. "Sequence convergence in the peptide-binding region of primate and rodent MHC class Ib molecules". In: *Molecular Biology and Evolution* 14 (1997), pp. 1035–1041. ISSN: 0737-4038.

[277] Ning Yu et al. "Nucleotide diversity in gorillas". In: *Genetics* 166.3 (2004), pp. 1375–1383. ISSN: 0016-6731. DOI: 10.1534/genetics.166.3.1375.

[278] Jianzhi Zhang and Sudhir Kumar. "Detection of convergent and parallel evolution at the amino acid sequence level". In: *Molecular Biology and Evolution* 14.5 (1997), pp. 527–36. ISSN: 0737-4038. DOI: 10.1093/oxfordjournals.molbev.a025789.

[279] Jianzhi Zhang and Masatoshi Nei. "Accuracies of ancestral amino acid sequences inferred by the parsimony, likelihood, and distance methods". In: *Journal of Molecular Evolution* 44.S1 (1997), S139–S146. ISSN: 0022-2844. DOI: 10.1007/PL00000067.

[280] Jianzhi Zhang, Rasmus Nielsen, and Ziheng Yang. "Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level". In: *Molecular Biology and Evolution* 22 (2005), pp. 2472–2479. ISSN: 07374038. DOI: 10.1093/molbev/msi237.

[281] Ying-Xin Zhang et al. "Genome shuffling leads to rapid phenotypic improvement in bacteria". In: *Nature* 415.6872 (2002), pp. 644–646. ISSN: 00280836. DOI: 10.1038/415644a.

[282] Xiaowei Zhuang et al. "Correlating structural dynamics and function in single ribozyme molecules". In: *Science* 296.5572 (2002), pp. 1473–1476. ISSN: 00368075. DOI: 10.1126/science.1069013.

[283] Michael Zuker and Patrick Stiegler. "Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information". In: *Nucleic Acids Research* 9.1 (1981), pp. 133–48. ISSN: 0305-1048.

[284] Hadas Zur and Tamir Tuller. "Strong association between mRNA folding strength and protein abundance in *S. cerevisiae*". In: *EMBO Reports* 13.3 (2012), pp. 272–7. ISSN: 1469-3178. DOI: 10.1038/embor.2011.262.

# Curriculum Vitae

**Surname:** REZAEE VAHDATI
**First name:** Reza Ali

**Date of birth:** 22.01.1989

**Nationality:** Iran

## Education

2012-2017 **University of Zurich, Zurich, Switzerland**
Institute of Evolutionary Biology and Environmental Studies
PhD candidate (Evolutionary Biology)
Employed as a PhD student since January 2012.

2010-2011 **University of Manchester, Manchester, UK**

MSc Evolutionary Genetics and Genomics
MSc Thesis 1: Identifying QTL and Genomic imprinting effects on mice tissue phenotypes.
MSc Thesis 2: Analysis of replication timing on genetic variation in the human genome.

2006-2010 **Azad University, Tonekabon branch, Iran**

BSc Cellular and Molecular Biology, Genetics

2002-2006 **Malek Ashtar Highschool, Mashhad, Iran**